A few notes on Statistical Learning Theory

Shahar Mendelson[1]

Contents

[1]Computer Sciences Laboratory, RSISE, The Australian National University, Canberra 0200, Australia. Email: shahar.mendelson@anu.edu.au

1 Introduction

In these notes our aim is to survey recent (and not so recent) results regarding the mathematical foundations of learning theory. The focus in this article is on the theoretical side and not on the applicative one; hence, we shall not present examples which may be interesting from the practical point of view but have little theoretical significance. This survey is far from being complete and it focuses on problems the author finds interesting (an opinion which is not necessarily shared by the majority of the learning community). Relevant books which present a more evenly balanced approach are, for example [1, 4, 35, 36]

The starting point of our discussion is the formulation of the learning problem. Consider a class G, consisting of real valued functions defined on a space Ω, and assume that each $g \in G$ maps Ω into $[0,1]$. Let T be an unknown function, $T : \Omega \to [0,1]$ and set μ to be an unknown probability measure on Ω.

The data one receives are a finite sample $(X_i)_{i=1}^n$, where (X_i) are independent random variables distributed according to μ, and the values of the unknown function on the sample $\big(T(X_i)\big)_{i=1}^n$. The objective of the learner is to construct a function in G which is almost the closest function to T in the set, with respect to the $L_2(\mu)$ norm. In other words, given $\varepsilon > 0$, one seeks a function $g_0 \in G$ which satisfies that

$$\mathbb{E}_\mu |g_0 - T|^2 \leq \inf_{g \in G} \mathbb{E}_\mu |g - T|^2 + \varepsilon, \tag{1.1}$$

where $\mathbb{E}_\mu$ is the expectation with respect to the probability measure μ. Of course, this function has to be constructed according to the data at hand.

A mapping L is a learning rule if it maps every sample $s_n = \big((X_i)_{i=1}^n, \big(T(X_i)\big)_{i=1}^n\big)$ to some $L_{s_n} \in G$. The measure of the effectiveness of the learning rule is "how much data" it needs in order to produce an almost optimal function in the sense of (1.1).

The one learning rule which seems to be the most natural (and it is the one we focus on throughout this article) is the loss minimization. For the sake of simplicity, we assume that the $L_2(\mu)$ minimal distance between T and members of G is attained at a point we denote by $P_G T$, and define a new function class, which is based on G and T in the following manner; for every $g \in G$, let $\ell(g) = |g - T|^2 - |P_G T - T|^2$ and set $\mathcal{L} = \{\ell(g) | g \in G\}$. $\mathcal{L}$ is called the 2-loss class associated with G and T, and there are obvious generalizations of this notion when other norms are considered.

For every sample $s_n = \{x_1, ..., x_n\}$ and $\varepsilon > 0$, let $g^* \in G$ be any function which satisfies that

$$\frac{1}{n} \sum_{i=1}^n \big(g^*(x_i) - T(x_i)\big)^2 \leq \inf_{g \in G} \frac{1}{n} \sum_{i=1}^n \big(g(x_i) - T(x_i)\big)^2 + \varepsilon. \tag{1.2}$$

Thus, any g^* is an "almost minimizer" of the *empirical distance* between members of G and the target T. To simplify the presentation, let us introduce a notation we shall use throughout these notes. Given a set $\{x_1, ..., x_n\}$, let μ_n be the empirical measure supported on the set. In other words, $\mu_n = n^{-1} \sum_{i=1}^n \delta_{x_i}$ where δ_{x_i} is the point evaluation functional on the set $\{x_i\}$. The $L_2(\mu_n)$ norm is defined as $\|f\|_{L_2(\mu_n)}^2 = n^{-1} \sum_{i=1}^n f^2(x_i)$. Therefore, g^* is defined as a function which satisfies that

$$\|g^* - T\|_{L_2(\mu_n)}^2 \leq \inf_{g \in G} \|g - T\|_{L_2(\mu_n)}^2 + \varepsilon.$$

From the definition of the loss class it follows that $\mathbb{E}_{\mu_n} \ell(g^*) \leq \varepsilon$. Indeed, the second term in every loss function is the same - $|T - P_G T|^2$, hence the infimum is determined only by the first term $|g - T|^2$. Thus,

$$\mathbb{E}_{\mu_n} \ell(g^*) \leq \inf_{f \in \mathcal{L}} \mathbb{E}_{\mu_n} f + \varepsilon \leq \varepsilon,$$

since $\inf_{f \in \mathcal{L}} \mathbb{E}_{\mu_n} f \leq 0$, simply by looking at $f = \ell(P_G T)$.

The question we wish to address is when such a function $\ell(g^*)$ will also be an "almost minimizer" with respect to the original L_2 norm. Since $\|g - T\|_{L_2(\mu)} \geq \|P_G T - T\|_{L_2(\mu)}$ it follows that for every $g \in G$, $\mathbb{E}_\mu \ell(g) \geq 0$. Therefore, our question is when

$$\mathbb{E}_\mu \ell(g^*) \leq \inf_{g \in G} \mathbb{E}_\mu \ell(g) + \varepsilon = \varepsilon? \tag{1.3}$$

Formally, we attempt to solve the following

Question 1.1 *Fix $\varepsilon > 0$, let s_n be a sample and set g^* to be a function which satisfies* (1.2). *Does it mean that $\mathbb{E}_\mu \ell(g^*) \leq \varepsilon$?*

Of course, it is too much to hope for that the answer is affirmative for any given sample, or even for any "long enough" sample, because one can encounter arbitrarily long samples that give misleading information on the behaviour of T. The hope is that an affirmative answer will be true with a relatively high probability as the size of the sample increases. The tradeoff between the desired accuracy ε, the high probability required and the size of the sample is the main question we wish to address.

Any attempt to approximate T with respect to any measure other than the measure according to which the sampling is made will not be successful. For example, if one has two probability measures which are supported on disjoint sets, any data received by sampling according to one measure will be meaningless when computing distances with respect to the other.

Another observation is that if the class G is "too large" it would be impossible to construct any worthwhile approximating function using empirical data. Indeed, assume that G consists of all the continuous functions on $[0, 1]$ which are bounded by 1, and for the sake of simplicity, assume that T is a Boolean function and that μ is the Lebesgue measure on $[0, 1]$. By a standard density argument, there are functions in G which are arbitrarily close to T with respect to the $L_2(\mu)$ distance, hence $\inf_{g\in G}\mathbb{E}_\mu|T-g|^2=0$. On the other hand, for any sample $\{(x_i),\big(T(x_i)\big)\}$ of T and every $\varepsilon>0$ there is some $g\in G$ which coincides with T on the sample, but $\mathbb{E}_\mu|T-g|^2\geq 1-\varepsilon$.

The problem one encounters in this example occurs because the class in question is too large; even if one receives as data an arbitrarily large sample, there are still "too many" very different functions in the class which behave in a similar way to (or even coincide with) T on the sample, but they are very far apart. In other words, if one wants an effective learning scheme, the structure of the class should not be too rich, in the sense that additional empirical data (i.e. a larger sample) decreases the number of class members which are "close" to the target on the data. Hence, all the functions which the learning algorithm may select become "closer" to the target as the size of the sample increases.

The two main approaches we focus on are outcomes of this line of reasoning. Firstly, assume that one can ensure that when the sample size is large enough, then with high probability, empirical means of members of $\mathcal{L}$ are uniformly close to the actual means (that is, with high probability every $f\in\mathcal{L}$ satisfies that, $|\mathbb{E}_\mu f-\mathbb{E}_{\mu_n}f|<\varepsilon$). In particular, if $\mathbb{E}_{\mu_n}\ell(g^*)<\varepsilon$ then $\mathbb{E}_\mu\ell(g^*)<2\varepsilon$. This naturally leads us to the definition of *Glivenko-Cantelli* classes.

Definition 1.2 *Let F be a class of functions. We say that F is a uniform Glivenko Cantelli class if for every $\varepsilon>0$,*

$$\lim_{n\to\infty}\sup_\mu Pr\Big\{\sup_{f\in F}|\mathbb{E}_\mu f-\frac{1}{n}\sum_{i=1}^n f(X_i)|\geq\varepsilon\Big\}=0,$$

where $(X_i)_{i=1}^\infty$ are independent random variables distributed according to μ.

The fact that the supremum is taken with respect to all probability measures μ is important because one does not have a-priori information on the probability measure according to which the data is sampled.

This definition has a quantified version. For every $0<\varepsilon,\delta<1$, let $S_F(\varepsilon,\delta)$ be the first integer n_0 such that for every $n\geq n_0$ and any probability measure μ,

$$Pr\Big\{\sup_{f\in F}|\mathbb{E}_\mu f-\mathbb{E}_{\mu_n}f|\geq\varepsilon\Big\}\leq\delta, \tag{1.4}$$

where μ_n is the random empirical measure $n^{-1}\sum_{i=1}^n\delta_{X_i}$.

S_F is called the Glivenko-Cantelli sample complexity of the class F with accuracy ε and confidence δ.

Of course, the ability to control the means of *every* function within the class is a very strong property, and is only a (loose!) sufficient condition which suffices to ensure that g^* is a "good approximation" of T. In fact, all that we are interested in is that this type of a condition holds for a function like $\ell(g^*)$ (i.e., an almost minimizer of $\ell(g)$ with respect to an empirical norm). Therefore, one would like to estimate

$$\sup_{\mu} Pr\Big\{\exists f \in \mathcal{L},\ \mathbb{E}_{\mu_n} f < \varepsilon, \mathbb{E}_{\mu} f \geq 2\varepsilon\Big\}. \tag{1.5}$$

Let $C_{\mathcal{L}}(\varepsilon, \delta)$ be the first integer such that for every $n \geq C_{\mathcal{L}}(\varepsilon, \delta)$ the term in (1.5) is smaller than δ. For such a value of n, there is a set of large probability on which any function which is an "almost minimizer" of the empirical loss will be an "almost minimizer" of the actual loss regardless of the underlying probability measure, implying that our learning algorithm will be successful.

These notes are divided into two main parts. The first one deals with Glivenko-Cantelli classes and the parameters which govern the sample complexity of such classes. In the second section we focus on (1.5) and show that under mild structural conditions on the class G it is possible to improve the estimates obtained using a Glivenko-Cantelli argument.

Notational conventions we shall use are that all absolute constants are denoted by c and C. Their values may change from line to line, or even within the same line. If X and Y are random variables, $\mathbb{E}f(X,Y)$ denotes the expectation with respect to both variables. The expectation with respect to X is denoted by $\mathbb{E}_X f(X,Y) = \mathbb{E}\big(f(X,Y)|Y\big)$.

2 Glivenko-Cantelli Classes

In this section we study the properties of uniform Glivenko-Cantelli classes (uGC classes for brevity), which are classes that satisfy (1.3) or (1.4). We examine various characterization theorems for uGC classes. The results which are relevant to the problem of sample complexity estimates are presented in full. We assume that the reader has some knowledge of the basic definitions in probability theory and empirical processes theory. One can turn to [5] for a more detailed introduction, or to [34, 8] for a complete and rigorous analysis.

We start this section with a presentation of the classical approach, using which sample complexity estimates for uGC classes were established in the past [37, 2]. This approach has its own merit, though the estimates one obtains using this method are suboptimal.

2.1 The classical approach

Let F be a class of functions whose range is contained in $[-1,1]$. We say that $(Z_i)_{i\in I}$ is a random process indexed by F if for every $f \in F$ and every $i \in I$, $Z_i(f)$ is a random variable. The process is called i.i.d. if the finite dimensional marginal distributions $\big(Z_i(f_1), ..., Z_i(f_k)\big)$ are independent random vectors[2].

[2]throughout these notes we are going to omit all the measurability issues one should address in a completely rigorous exposition.

One example the reader should have in mind is the following random process: let μ be a probability measure on the domain Ω and let $X_1, ..., X_n$ be independent random variables distributed according to μ. Set μ_n to be the empirical measure supported on $X_1, ..., X_n$ - which is $n^{-1}\sum_{i=1}^n \delta_{X_i}$. Hence, μ_n is a *random* probability measure given by the average of point masses at X_i. Let $Z_i(\cdot) = \big(\delta_{X_i} - \mu\big)(\cdot)$, where the last equation should be interpreted as $Z_i(f) = f(X_i) - \mathbb{E}_\mu(f)$ for every $f \in F$. Note that $Z_1, ..., Z_n$ is an i.i.d. process with 0 mean (since for every $f \in F$, $\mathbb{E}Z_i(f) = 0$). Moreover,

$$\sup_{f\in F} |\sum_{i=1}^n Z_i(f)| = \sup_{f \in F} |\sum_{i=1}^n \big(f(X_i) - \mathbb{E}_\mu f\big)|,$$

which is exactly the random variable we are interested in.

Our strategy is based on the following idea, which, for the sake of simplicity, is explained for the trivial class consisting of a single element. We wish to measure "how close" empirical means are to the actual mean. If this holds with high probability, then two random empirical means should be "close" to each other. Thus, if (X_i') is an independent copy of (X_i), then the probability that $|\sum_{i=1}^n \big(f(X_i) - f(X_i')\big)| \geq x$ should be an indication of the probability of deviation of the empirical means from the actual one. By symmetry, for every i, $Y_i = f(X_i) - f(X_i')$ is distributed as $-Y_i$. Hence, for every selection of signs ε_i,

$$Pr\big\{\big|\sum_{i=1}^n f(X_i) - f(X_i')\big| \geq x\big\} = Pr\big\{\big|\sum_{i=1}^n \varepsilon_i\big(f(X_i) - f(X_i')\big)\big| \geq x\big\}. \tag{2.1}$$

Now, consider $(\varepsilon_i)_{i=1}^n$ as independent Rademacher (i.e. symmetric $\{-1,1\}$-valued) random variables, and (2.1) still holds, where Pr on the right hand side now denotes the product measure generated by (X_i), (X_i') and (ε_i). By the triangle inequality,

$$\begin{aligned} Pr\big\{\big|\sum_{i=1}^n f(X_i) - f(X_i')\big| \geq x\big\} &\leq Pr\big\{\big|\sum_{i=1}^n \varepsilon_i f(X_i)\big| \geq \frac{x}{2}\big\} + Pr\big\{\big|\sum_{i=1}^n \varepsilon_i f(X_i')\big| \geq \frac{x}{2}\big\} \\ &= 2Pr\big\{\big|\sum_{i=1}^n \varepsilon_i f(X_i)\big| \geq \frac{x}{2}\big\} \end{aligned}$$

since X_i and X_i' are identically distributed. Therefore, $Pr\{|\sum_{i=1}^n \varepsilon_i f(X_i)| \geq x/2\}$ could be the right quantity to control the deviation we require.

Since this is far from being rigorous, one has to make the above reasoning precise. There are two main issues that need to be resolved; firstly, can this kind of a result be true for a "rich" class of functions - consisting of more than a single function, and secondly, how can one control the probability of deviation even after this "symmetrization" argument?

2.1.1 The symmetrization procedure

Here we present the following symmetrization argument, due to Giné and Zinn [10], which is the first step in the "classical" approach.

Theorem 2.1 *Let $(Z_i)_{i=1}^n$ be an i.i.d. stochastic process which has* 0 *mean, and for every* $1 \leq i \leq n$, *set* $h_i : F \to \mathbb{R}$ *to be an arbitrary function. Then, for every* $x > 0$

$$\Big(1 - \frac{4n}{x^2} \sup_{f \in F} \operatorname{var}\big(Z_1(f)\big)\Big) Pr\Big\{\sup_{f \in F} |\sum_{i=1}^n Z_i(f)| > x\Big\}$$
$$\leq 2Pr\big\{\sup_{f \in F} |\sum_{i=1}^n \varepsilon_i\big(Z_i(f) - h_i(f)\big)| > \frac{x}{4}\big\},$$

where $(\varepsilon_i)_{i=1}^n$ *are independent Rademacher random variables.*

Before proving this theorem, let us consider its implications for "our" empirical process. Fix a probability measure μ according to which the sampling is made. Then, $Z_i(f) = f(X_i) - \mathbb{E}_\mu f$ and put $h_i(f) = -\mathbb{E}_\mu f$. Also, set $v^2 = \sup_{f \in F} \operatorname{var}(f)$, and note that if $x \geq 2\sqrt{2}\sqrt{n}v$ then $1 - \frac{4n}{x^2} \sup_{f \in F} \operatorname{var}\big(Z_1(f)\big) \geq 1/2$. Therefore, for such a value of x,

$$Pr\Big\{\sup_{f \in F} |\sum_{i=1}^n \big(f(X_i) - \mathbb{E}_\mu f\big)| > x\Big\} \leq 4Pr\Big\{\sup_{f \in F} |\sum_{i=1}^n \varepsilon_i f(X_i)| > \frac{x}{4}\Big\}. \tag{2.2}$$

Now, fix any $\varepsilon > 0$ and let $x = n\varepsilon$. If $n \geq 8v^2/\varepsilon^2$ then

$$Pr\Big\{\sup_{f \in F} |\frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}_\mu f| > \varepsilon\Big\} \leq 4Pr\Big\{\sup_{f \in F} |\sum_{i=1}^n \varepsilon_i f(X_i)| > \frac{n\varepsilon}{4}\Big\}. \tag{2.3}$$

In particular, if each function in F maps Ω into $[-1, 1]$ then $v^2 \leq 1$. Thus, (2.3) holds for any $n \geq 8/\varepsilon^2$.

Proof of theorem 2.1: Let W_i be an independent copy of Z_i and fix $x > 0$. Denote by P_Z (resp. P_W) the probability measure associated with the process (Z_i) (resp. (W_i)). Put $\beta = \inf_{f \in F} Pr\{|\sum_{i=1}^n Z_i(f)| < x/2\}$ and let $A = \{\sup_{f \in F} |\sum_{i=1}^n Z_i(f)| > x\}$. For every element in A there is a realization of the process Z_i and some $f \in F$ such that $|\sum_{i=1}^n Z_i(f)| > x$. Fix this realization and f and observe that by the triangle inequality, if $|\sum_{i=1}^n W_i(f)| < x/2$ then $|\sum_{i=1}^n Z_i(f) - W_i(f)| > x/2$. Since $(W_i)_{i=1}^n$ is a copy of $(Z_i)_{i=1}^n$ then

$$\beta \leq P_W\Big\{|\sum_{i=1}^n W_i(f)| < \frac{x}{2}\Big\} \leq P_W\Big\{|\sum_{i=1}^n W_i(f) - \sum_{i=1}^n Z_i(f)| > \frac{x}{2}\Big\}$$
$$\leq P_W\Big\{\sup_{f \in F} |\sum_{i=1}^n W_i(f) - \sum_{i=1}^n Z_i(f)| > \frac{x}{2}\Big\}.$$

Since the two extreme sides of this inequality are independent of the specific selection of f, this inequality holds on the set A. Integrating with respect to Z on A it follows that

$$\beta P_Z\Big\{\sup_{f\in F}|\sum_{i=1}^n Z_i(f)|>x\Big\}\leq P_ZP_W\Big\{\sup_{f\in F}|\sum_{i=1}^n\big(Z_i(f)-W_i(f)\big)|>\frac{x}{2}\Big\}.$$

Clearly, $Z_i - W_i$ has the same distribution as $W_i - Z_i = -(Z_i - W_i)$, implying that for every selection of signs $(\varepsilon_i)_{i=1}^n \in \{-1,1\}^n$, $\sum_{i=1}^n Z_i - W_i$ has the same distribution as $\sum_{i=1}^n \varepsilon_i(Z_i - W_i)$. Hence,

$$\begin{aligned}P_ZP_W\Big\{\sup_{f\in F}|\sum_{i=1}^n\big(Z_i(f)-W_i(f)\big)|>\frac{x}{2}\Big\}&=P_ZP_W\Big\{\sup_{f\in F}|\sum_{i=1}^n\varepsilon_i\big(Z_i(f)-W_i(f)\big)|>\frac{x}{2}\Big\}\\&=\mathbb{E}_\varepsilon P_ZP_W\Big\{\sup_{f\in F}|\sum_{i=1}^n\varepsilon_i\big(Z_i(f)-W_i(f)\big)|>\frac{x}{2}\Big\},\end{aligned}$$

where $\mathbb{E}_\varepsilon$ denotes the expectation with respect to the Rademacher random variables $(\varepsilon_i)_{i=1}^n$. By the triangle inequality, for every selection of functions h_i and every fixed realization $(\varepsilon_i)_{i=1}^n$,

$$P_ZP_W\Big\{\sup_{f\in F}|\sum_{i=1}^n\varepsilon_i\big(Z_i(f)-W_i(f)\big)|>\frac{x}{2}\Big\}\leq 2P_Z\Big\{\sup_{f\in F}|\sum_{i=1}^n\varepsilon_i\big(Z_i(f)-h_i(f)\big)|>\frac{x}{2}\Big\},$$

and by Fubini's Theorem

$$\mathbb{E}_\varepsilon\Big(P_Z\Big\{\sup_{f\in F}|\sum_{i=1}^n\varepsilon_i\big(Z_i(f)-h_i(f)\big)|>\frac{x}{2}\Big\}\Big|(\varepsilon_i)_{i=1}^n\Big)=Pr\Big\{\sup_{f\in F}|\sum_{i=1}^n\varepsilon_i\big(Z_i(f)-h_i(f)\big)|>\frac{x}{2}\Big\}.$$

Finally, to estimate β, note that by Chebyshev's inequality

$$Pr\Big\{|\sum_{i=1}^n Z_i(f)|>\frac{x}{2}\Big\}\leq\frac{4n}{x^2}\mathrm{var}\big(Z_1(f)\big),$$

for every $f \in F$, and thus, $\beta \geq 1-(4n/x^2)\sup_{f\in F}\mathrm{var}\big(Z_1(f)\big)$.

■

After establishing (2.3), the next step is to transform a very rich class to a trivial class, consisting of a single function, and then estimate $Pr\big\{\big|\sum_{i=1}^n\varepsilon_i f(x_i)\big|>x\big\}$. We show that one can effectively replace the (possibly) infinite class F with a finite set which approximates the original class in some sense. The "richness" of the class F will be reflected by the cardinality of the finite approximating set. This approximation scheme is commonly used in many areas of mathematics, and the main notion behind it is called *covering numbers*.

2.1.2 Covering numbers and complexity estimates

Let (Y,d) be a metric space and set $F \subset Y$. For every $\varepsilon > 0$, denote by $N(\varepsilon, F, d)$ the minimal number of open balls (with respect to the metric d) needed to cover F. That is, the minimal cardinality of the set $\{y_1, ..., y_m\} \subset Y$ with the property that every $f \in F$ has is some y_i such that $d(f, y_i) < \varepsilon$. The set $\{y_1, ..., y_m\}$ is called an ε-cover of F. The logarithm of the covering numbers is called the entropy of the set.

We will be interested in metrics endowed by samples; for every sample $\{x_1, ..., x_n\}$ let μ_n be the empirical measure supported on that sample. For $1 \leq p < \infty$ and a function f, put $\|f\|_{L_p(\mu_n)} = \big(n^{-1}\sum_{i=1}^n |f(x_i)|^p\big)^{1/p}$ and set $\|f\|_\infty = \max_{1\leq i\leq n}|f(x_i)|$. Let $N\big(\varepsilon, F, L_p(\mu_n)\big)$ be the covering numbers of F at scale ε with respect to the $L_p(\mu_n)$ norm.

Two easy observations we require are the following. Firstly, if $n^{-1}|\sum_{i=1}^n f(x_i)| > t$ and if $\|f-g\|_{L_1(\mu_n)} < t/2$ then

$$\frac{1}{n}|\sum_{i=1}^n g(x_i)| \geq \frac{1}{n}|\sum_{i=1}^n f(x_i)| - \frac{1}{n}\sum_{i=1}^n |f(x_i) - g(x_i)| > \frac{t}{2}.$$

Secondly, for every empirical measure μ_n and every $1 \leq p \leq \infty$, $\|f\|_{L_1(\mu_n)} \leq \|f\|_{L_p(\mu_n)} \leq \|f\|_{L_\infty(\mu_n)}$. Hence,

$$N\big(\varepsilon, F, L_1(\mu_n)\big) \leq N\big(\varepsilon, F, L_p(\mu_n)\big) \leq N\big(\varepsilon, F, L_\infty(\mu_n)\big).$$

In a similar fashion to the notion of covering numbers one can define the packing numbers of a class. Roughly speaking, a packing number is the maximal cardinality of a subset of F with the property that the distance between any two of its members is "large".

Definition 2.2 *Let (X,d) be a metric space. We say that $K \subset X$ is ε-separated with respect to the metric d if for every $k_1, k_2 \in K$, $d(k_1,k_2) \geq \varepsilon$.*

Given a set $F \subset X$, define its ε-packing number as the maximal cardinality of a subset of F which is ε-separated, and denote it by $D(\varepsilon, F, d)$.

It is easy to see that the covering numbers and the packing numbers are closely related. Indeed, assume that $K \subset F$ is a maximal ε-separated subset. By the maximality, for every $f \in F$ there is some $k \in K$ for which $d(x,k) < \varepsilon$, which shows that $N(\varepsilon, F, d) \leq D(\varepsilon, F, d)$. On the other hand, let $\{y_1, ..., y_m\}$ be an $\varepsilon/2$ cover of F and assume that $f_1, ..., f_k$ is a maximal ε-separated subset of F. In every ball $\{y|d(y,y_i) < \varepsilon/2\}$ there is at most a single element of the packing (by the triangle inequality, the diameter of this ball is smaller than ε). Since this is true for any cover of F then $D(\varepsilon, F, d) \leq N(\varepsilon/2, F, d)$.

Our discussion will rely heavily on covering and packing numbers. We can now combine the symmetrization argument with the notion of covering numbers and obtain the required complexity estimates.

Theorem 2.3 *Let F be a class of functions which map Ω into $[-1,1]$ and set μ to be a probability measure on Ω. Let $(X_i)_{i=1}^{\infty}$ be independent random variables distributed according to μ. For every $\varepsilon > 0$ and any $n \geq 8/\varepsilon^2$,*

$$Pr\Big\{\sup_{f \in F} |\frac{1}{n}\sum_{i=1} f(X_i) - \mathbb{E}_\mu f| > \varepsilon\Big\} \leq 8\mathbb{E}_\mu\big[N\big(\varepsilon, F, L_1(\mu_n)\big)\big]e^{-\frac{n\varepsilon^2}{128}},$$

where μ_n is the (random) empirical measure supported on $\{X_1, ..., X_n\}$.

One additional preliminary result we need before proceeding with the proof will enable us to handle the "trivial" case of classes consisting of a single function. This case follows from Hoeffding's inequality [12, 34]

Theorem 2.4 *Let $(a_i)_{i=1}^n \subset \mathbb{R}$ and let $(\varepsilon_i)_{i=1}^n$ be independent Rademacher random variables (that is, symmetric $\{-1,1\}$-valued). Then,*

$$Pr\Big\{|\sum_{i=1}^n \varepsilon_i a_i| > x\Big\} \leq 2e^{-\frac{1}{2}x^2/\|a\|_2},$$

where $\|a\|_2 = \big(\sum_{i=1}^n a_i^2\big)^{1/2}$.

In our case, $(a_i)_{i=1}^n$ is going to be the values of the function f on a fixed sample $\{x_1, ..., x_n\}$.

Proof of theorem 2.3: Let $A = \big\{\sup_{f\in F} |\sum_{i=1}^n \varepsilon_i f(X_i)| > \frac{n\varepsilon}{4}\big\}$, and denote by χ_A the characteristic function of A. By Fubini's Theorem,

$$Pr(A) = \mathbb{E}_\mu\big(\mathbb{E}_\varepsilon \chi_A | X_1, ..., X_n\big) = \mathbb{E}_\mu\big(Pr\{\sup_{f\in F} |\sum_{i=1}^n \varepsilon_i f(X_i)| > \frac{n\varepsilon}{4}\}|X_1, ..., X_n\big). \qquad (2.4)$$

Fix a realization of $X_1, ..., X_n$ and let μ_n be the empirical measure supported on that realization. Set G to be an $\varepsilon/8$ cover of F with respect to the $L_1(\mu_n)$ norm. Since F consists of functions which are bounded by 1, we can assume that the same holds for every $g \in G$. If $\sup_{f\in F} |\sum_{i=1}^n \varepsilon_i f(X_i)| > n\varepsilon/4$, there is some $f \in F$ for which this inequality holds. G is an $\varepsilon/8$-cover of F with respect to the $L_1(\mu_n)$, hence, there is some $g \in G$ which satisfies that $n^{-1}\sum_{i=1}^n |f(X_i) - g(X_i)| < \varepsilon/8$. Therefore, $\sup_{g\in G} |\sum_{i=1}^n \varepsilon_i g(X_i)| > n\varepsilon/8$, implying that for that realization of (X_i),

$$Pr\{\sup_{f\in F} |\sum_{i=1}^n \varepsilon_i f(X_i)| > \frac{n\varepsilon}{4}\} \leq Pr\{\sup_{g\in G} |\sum_{i=1}^n \varepsilon_i g(X_i)| > \frac{n\varepsilon}{8}\}.$$

Applying the union bound, Hoeffding's inequality and the fact that for every $g \in G$, $\sum_{i=1}^n g(x_i)^2 \leq n$,

$$Pr\{\sup_{g\in G} |\sum_{i=1}^n \varepsilon_i g(X_i)| > \frac{n\varepsilon}{8}\} \leq 2|G| Pr\{|\sum_{i=1}^n \varepsilon_i g(X_i)| > \frac{n\varepsilon}{8}\} \leq 2N\big(\frac{\varepsilon}{8}, F, L_1(\mu_n)\big)e^{-\frac{n\varepsilon^2}{128}}.$$

Finally, our claim follows from (2.4) and (2.3).

■

Unfortunately, it might be very difficult to compute the expectation of the covering numbers. Thus, one natural thing to do is to introduce *uniform entropy numbers.*

Definition 2.5 *For every class F, $1 \leq p \leq \infty$ and $\varepsilon > 0$, let*

$$\log N_p\big(\varepsilon, F, n\big) = \sup_{\mu_n} \log N\big(\varepsilon, F, L_p(\mu_n)\big),$$

and

$$\log N_p\big(\varepsilon, F\big) = \sup_n \sup_{\mu_n} \log N\big(\varepsilon, F, L_p(\mu_n)\big).$$

$\log N_p(\varepsilon, F)$ *are called the uniform entropy numbers of F with respect to the $L_p(\mu_n)$.*

The only hope for establishing non-trivial uniform entropy bounds is when the covering numbers do not depend on the cardinality of the set on which the empirical measure is supported. In some sense, this implies that classes for which one can obtain uniform entropy bounds must be "small". As we will show in sections to come, one can establish such dimension-free bounds in terms of the combinatorial parameters which are used to "measure" the size of a class of functions.

The following result seems to be a weaker version of the theorem, but in the sequel we prove that it is a necessary condition for the uniform GC property as well.

Theorem 2.6 *Assume that F is a class of functions which are all bounded by 1. If there is some $1 \leq p \leq \infty$ such that for every $\varepsilon > 0$ the uniform entropy numbers satisfy*

$$\lim_{n \to \infty} \frac{\log N_p(\varepsilon, F, n)}{n} = 0,$$

then F is a uniform Glivenko-Cantelli class.

An easy observation is that it is possible to bound the Glivenko-Cantelli sample complexity using the uniform entropy numbers of the class.

Theorem 2.7 *Let F be a class of functions which map Ω into $[-1, 1]$. Then for every $0 < \varepsilon, \delta < 1$,*

$$\sup_\mu Pr\Big\{\sup_{f \in F} \Big|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_\mu f\Big| \geq \varepsilon\Big\} \leq \delta,$$

provided that $n \geq \frac{128}{\varepsilon^2}\big(\log N_1(\varepsilon, F) + \log(8/\delta)\big)$.

In particular, if the uniform entropy is of power type q (that is, $\log N_1(\varepsilon, F) = O(\varepsilon^{-q})$), then the uGC sample complexity is (up to logarithmic factors in δ^{-1}) $O(\varepsilon^{-(2+q)})$.

As an example, assume that F is the 2-loss class associated with G and T. In this case, the L_p entropy numbers of the loss class can be controlled by those of G.

Lemma 2.8 *Let G be a class of functions whose range is contained in $[0,1]$ and assume that the same holds for T. If $\mathcal{L}$ is the 2-loss class associated with G and T, then for every $\varepsilon > 0$, every $1 \le p \le \infty$ and every probability measure μ,*

$$N\big(\varepsilon, \mathcal{L}, L_p(\mu)\big) \le N\big(\frac{\varepsilon}{4}, G, L_p(\mu)\big).$$

Proof: Since $\mathcal{L}$ is a shift of the class $(G-T)^2$, and since covering numbers of a shifted class are the same as those of the original one (a shift is an isometry with respect to the L_p norm), it is enough to estimate the covering numbers of the class $(G-T)^2$. Let $\{y_1, ..., y_m\}$ be an ε-cover of G in $L_p(\mu)$. If $\|g - y_i\|_{L_p(\mu)} < \varepsilon$, then pointwise

$$|g-T|^2 - |y_i - T|^2 = |g - y_i| \cdot |g + y_i - 2T| \le 4|g - y_i|.$$

Hence, $\| |g-T|^2 - |y_i - T|^2 \|_{L_p(\mu)} \le 4\|g - y_i\|_{L_p(\mu)} < 4\varepsilon$.

■

Corollary 2.9 *Using the notation of the previous theorem, for every $0 < \varepsilon, \delta < 1$,*

$$S_{\mathcal{L}}(\varepsilon, \delta) \le \frac{128}{\varepsilon^2}\big(\log N_1\big(\varepsilon/4, G\big) + \log(8/\delta)\big)$$

The natural question which comes to mind is how to estimate the uniform entropy numbers of a class. Historically, this was the reason for the introduction of several combinatorial parameters. We will show that by using them one can control the uniform entropy.

2.2 Combinatorial parameters and covering numbers

The first combinatorial parameter was introduced by Vapnik and Chervonenkis [37] to control the empirical L_∞ entropy of Boolean classes of functions.

Definition 2.10 *Let F be a class of $\{0,1\}$-valued functions on a space Ω. We say that F shatters $\{x_1, ..., x_n\} \subset \Omega$, if for every $I \subset \{1, ..., n\}$ there is a function $f_I \in F$ for which $f_I(x_i) = 1$ if $i \in I$ and $f_I(x_i) = 0$ if $i \notin I$. Let*

$$VC(F, \Omega) = \sup\Big\{|A| \,\Big|\, A \subset \Omega,\ A \text{ is shattered by } F\Big\}.$$

$VC(F, \Omega)$ is called the VC dimension of F, but when the underlying space is clear we denote it by $VC(F)$.

The VC dimension has a geometric interpretation. A set $s_n = \{x_1, .., x_n\}$ is shattered if the set $\{\big(f(x_1), ..., f(x_n)\big)|f \in F\}$ is the combinatorial cube $\{0,1\}^n$. For every sample σ denote by $P_\sigma F$ the coordinate projection of F,

$$P_\sigma F = \{\big(f(x_i)\big)_{x_i \in \sigma} \big| \ f \in F\}.$$

Hence, the VC dimension is the largest cardinality of $\sigma \subset \Omega$ such that $P_\sigma F$ is the combinatorial cube of dimension $|\sigma|$.

Next, we present bounds on the empirical L_∞ and L_2 uniform entropy estimate using the VC dimension.

2.2.1 Uniform entropy and the VC dimension

We begin with the L_∞ estimates mainly for historical reasons. The following lemma, known as the Sauer-Shelah Lemma was proved independently at least three times, by Sauer [29], Shelah [30] and Vapnik and Chervonenkis [37].

Lemma 2.11 *Let F be a class of Boolean functions and set $d = VC(F)$. Then, for every finite subset $\sigma \subset \Omega$ of cardinality n,*

$$|P_\sigma F| \leq \Big(\frac{en}{d}\Big)^d.$$

In particular, for every $\varepsilon > 0$, $N\big(\varepsilon, F, L_\infty(\sigma)\big) \leq |P_\sigma F| \leq \big(en/d\big)^d$.

Using the Sauer-Shelah Lemma, one can characterize the uniform Glivenko-Cantelli property of a class of Boolean functions in terms of the VC dimension.

Theorem 2.12 *Let F be a class of Boolean functions. Then F is a uniform Glivenko-Cantelli class if and only if it has a finite VC dimension.*

Proof: Assume that $\text{VC}(F) = \infty$ and fix an integer $d \geq 2$. There is a set $\sigma \subset \Omega$, $|\sigma| = d$ such that $P_\sigma F = \{0,1\}^d$, and let μ be the uniform measure on σ (assigns a weight of $1/d$ to every point). For any $A \subset \sigma$ of cardinality $n \leq d/2$, let μ_n^A be the empirical measure supported on A. Since there is some $f_A \in F$ which is 1 on A and vanishes on $\sigma \backslash A$ then $|\mathbb{E}_\mu f_A - \mathbb{E}_{\mu_n} f_A| = |1 - n/d| \geq 1/2$. Hence, $\sup_{f \in F} |\mathbb{E}_{\mu_n^A} f - \mathbb{E}_\mu f| \geq 1/2$. Therefore, for any $n \leq d/2$,

$$Pr\Big\{\sup_{f \in F} |\mathbb{E}_{\mu_n} f - \mathbb{E}_\mu f| \geq 1/2\Big\} = 1,$$

and since d can be made arbitrarily large, F is not a uniform GC class.

To prove the converse, recall that for every $0 < \varepsilon < 1$ and every empirical measure μ_n supported on the sample s_n, $N\big(\varepsilon, F, L_\infty(s_n)\big) \leq |P_{s_n} F| \leq \big(en/d\big)^d$. Since the empirical

L_1 entropy is bounded by the empirical L_∞ one, $\log N_1(\varepsilon, F, n) \leq d \log(en/d)$. Thus, for every $\varepsilon > 0$, $\log N_1(\varepsilon, F, n) = o(n)$, implying that F is a uniform GC class.

■

In a similar fashion one can characterize the uGC property for Boolean classes using the L_p entropy numbers.

Corollary 2.13 *Let F be a class of Boolean functions. Then, F is a uniform Glivenko-Cantelli class if and only if for every $1 \leq p \leq \infty$ and every $\varepsilon > 0$, $\log N_p(\varepsilon, F, n) = o(n)$.*

Proof: Fix any $1 \leq p \leq \infty$. If for every $\varepsilon > 0$ $\log N_p(\varepsilon, F, n) = o(n)$, then by theorem 2.3, F is a uGC class. Conversely, if F is a uGC class then it has a finite VC dimension. Denote $VC(F) = d$, let σ be a sample of cardinality n and set μ_n to be the empirical measure supported on σ. For every $\varepsilon > 0$ and $1 \leq p < \infty$

$$\log N\big(\varepsilon, F, L_p(\mu_n)\big) \leq \log N\big(\varepsilon, F, L_\infty(\sigma)\big) \leq \log |P_\sigma F| \leq d \log\big(\frac{en}{d}\big) = o(n).$$

■

There is some hope that with respect to a "weaker" norm, one will be able to obtain uniform entropy estimates (which can not be derived from the L_∞ bounds presented here), that would lead to improved complexity bounds. Although the uGC property is characterized by the entropy with respect to any L_p norm (and in that sense, the L_∞ one is as good as any other L_p norm), from the quantitative point of view, it is much more desirable to obtain L_1 or L_2 entropy estimates, which will prove to be considerably smaller than the L_∞ ones.

Therefore, the next order of business is to estimate the uniform entropy of a VC class with respect to empirical L_p norms. This result is due to Dudley [7] and it is based on a combination of an extraction principle and the Sauer-Shelah Lemma. The probabilistic extraction argument simply states that if $K \subset F$ is "well separated" in $L_1(\mu_n)$ in the sense that every two points are different on a number of coordinated which is proportional to n, one can find a much smaller set of coordinates (which depends of the cardinality of K) on which every two points in K are different on at least one coordinate.

Theorem 2.14 *Let F be a class of Boolean functions and assume that $VC(F) = d$. Then, for every $1 \leq p < \infty$,*

$$N_p(\varepsilon, F) \leq \Big((2pe^2)\log\frac{2e^2}{\varepsilon}\Big)^d \Big(\frac{1}{\varepsilon}\Big)^{pd}.$$

Proof: Since the functions in F are $\{0,1\}$-valued, it is enough to prove the claim for $p = 1$. The general case follows since for any $f, g \in F$ and any probability measure μ, $\|f - g\|_{L_p(\mu)}^p = \|f - g\|_{L_1(\mu)}$.

Let $\mu_n = n^{-1}\sum_{i=1}^n \delta_{x_i}$ and fix $0 < \varepsilon < 1$. Set K_ε to be any ε-separated subset of F with respect to the $L_1(\mu_n)$ norm and denote its cardinality by D.

If $V = \{f_i - f_j | f_i \neq f_j \in K_\varepsilon\}$ then every $v \in V$ has at least $n\varepsilon$ coordinates which belong to $\{-1,1\}$. Indeed, since K_ε is ε-separated then for any $v \in V$

$$\varepsilon \leq \|v\|_{L_1(\mu_n)} = \|f_i - f_j\|_{L_1(\mu_n)} = \frac{1}{n}\sum_{l=1}^{n} |f_i(x_l) - f_j(x_l)| = \frac{1}{n}\sum_{l=1}^{n} |v(x_l)|,$$

and for every $1 \leq l \leq n$, $|v(x_l)| = |f_i(x_l) - f_j(x_l)| \in \{0,1\}$. In addition, it is easy to see that $|V| \leq D^2$.

Take $(X_i)_{i=1}^t$ to be independent $\{x_1, ..., x_n\}$-valued random variables, such that for every $1 \leq i \leq t$ and $1 \leq j \leq n$, $Pr(X_i = x_j) = 1/n$. It follows that for any $v \in V$,

$$Pr\Big\{\forall i,\ v(X_i) = 0\Big\} = \prod_{i=1}^{t} Pr\big\{v(X_i) = 0\big\} \leq (1-\varepsilon)^t.$$

Hence,

$$P\Big\{\exists v \in V,\ \forall i,\ v(X_i) = 0\Big\} \leq |V|\,(1-\varepsilon)^t \leq D^2(1-\varepsilon)^t.$$

Therefore,

$$P\Big\{\forall v \in V,\ \exists i,\ 1 \leq i \leq t\ |v(X_i)| = 1\Big\} \geq 1 - D^2(1-\varepsilon)^t,$$

and if the latter is greater than 0, there is a set of $\sigma \subset \{1, ..., n\}$ such that $|\sigma| = t$ and

$$|P_I K_\varepsilon| = \Big|\big\{\big(f(x_i)\big)_{i\in\sigma} \big| f \in K_\varepsilon\big\}\Big| = D.$$

Select $t = \frac{2\log D}{\varepsilon}$ which suffices to ensure the existence of such a set σ. By the Sauer-Shelah Lemma,

$$D = |P_\sigma K_\varepsilon| \leq |P_\sigma F| \leq \big(\frac{e|\sigma|}{d}\big)^d = \big(\frac{2e\log D}{d\varepsilon}\big)^d. \tag{2.5}$$

It is easy to see that if $\alpha \geq 1$ and $\alpha \log^{-1}\alpha \leq \beta$ then

$$\alpha \leq \beta \log(e\beta\log\beta).$$

Applying this to (2.5),

$$\log D \leq d\log\Big(\frac{2e^2}{\varepsilon}\log(\frac{2e}{\varepsilon})\Big),$$

as claimed.

■

This result was strengthened by Haussler in [11] in a very difficult proof, which removed the superfluous logarithmic factor.

Theorem 2.15 *There is an absolute constant C which satisfies that for every Boolean class F, any $1 \leq p < \infty$ and every $\varepsilon > 0$, $N_p(\varepsilon, F) \leq Cd(4e)^d\varepsilon^{-pd}$, where* $\mathrm{VC}(F) = d$.

The significance of theorem 2.14 and theorem 2.15 is that they provide uniform L_p entropy estimates for VC classes, while the L_∞ estimates are not dimension-free. These uniform entropy bounds play a very important role in our discussion. In particular, they can be used to obtain uGC complexity estimated for VC classes, using theorem 2.3.

Theorem 2.16 *Let F be a class of Boolean functions which has a finite VC dimension d. Then, for every $0 < \varepsilon, \delta < 1$,*

$$\sup_\mu Pr\Big\{\sup_{f \in F} |\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_\mu f| \geq \varepsilon\Big\} \leq \delta,$$

provided that $n \geq \frac{128}{\varepsilon^2}\big(d\log(2e^2/\varepsilon) + \log(8/\delta)\big)$.

Using the same reasoning and by lemma 2.8 it is possible to prove analogous results when F is the 2-loss class associated with a VC class and an arbitrary target T which maps Ω into $[0,1]$.

2.2.2 Generalized combinatorial parameters

After obtaining covering results (and generalization bounds) in the Boolean case, we attempt to extend our analysis to classes of real-valued functions. We focus on classes which consist of uniformly bounded functions, though it is possible to obtain some results in a slightly more general scenario ([34]). Hence, throughout this section F will denote a class of functions which map Ω into $[-1,1]$.

The path we take here is very similar to the one we used for VC classes. Firstly, one has to define a combinatorial parameter which measures the "size" of the class.

Definition 2.17 *For every $\varepsilon > 0$, a set $\sigma = \{x_1, ..., x_n\} \subset \Omega$ is said to be ε-shattered by F if there is some function $s : \sigma \to \mathbb{R}$, such that for every $I \subset \{1, ..., n\}$ there is some $f_I \in F$ for which $f_I(x_i) \geq s(x_i) + \varepsilon$ if $i \in I$, and $f_I(x_i) \leq s(x_i) - \varepsilon$ if $i \notin I$. Let*

$$\mathrm{fat}_\varepsilon(F) = \sup\Big\{|\sigma| \Big| \sigma \subset \Omega,\ \sigma \text{ is } \varepsilon\text{−shattered by } F\Big\}.$$

f_I is called the shattering function of the set I and the set $\big\{s(x_i)|x_i \in \sigma\big\}$ is called a witness to the ε-shattering.

The first bounds on the empirical L_∞ covering numbers in terms of the fat-shattering dimension was established in [2], where is was shown that F is a uGC class if and only if it has a finite fat-shattering dimension for every ε. The proof that if F is a uGC it has a finite fat-shattering dimension for every ε follows from a similar argument to the one used in the VC case. For the converse one requires empirical L_∞ entropy estimates combined with theorem 2.6. Dimension-free L_p entropy results for $1 \leq p < \infty$ in terms of the fat-shattering dimension were first proved in [19]. Both these results were improved in [22] and then in [23]. The proofs of all the results mentioned here are very difficult, and go beyond the scope of these notes. The second part of the following claim is due to Vershynin (still unpublished).

Theorem 2.18 *There is an absolute constant C which satisfies that for every $F \subset B\big(L_\infty(\Omega)\big)$, every sample s_n, every $1 \leq p < \infty$ and any $0 < \varepsilon < 1$,*

$$N\big(\varepsilon, F, L_p(\mu)\big) \leq \Big(\frac{2}{\varepsilon}\Big)^{K_p \mathrm{fat}_{c_p\varepsilon}(F)},$$

and

$$\log N\big(\varepsilon, F/s_n, L_\infty(s_n)\big) \leq K \cdot \mathrm{fat}_{c\varepsilon}(F) \log^{1+\delta}\Big(\frac{n}{\delta\varepsilon}\Big),$$

where K_p and c_p are constants which depend only on p.

The significance of these entropy estimates goes far beyond learning theory. They are essential in solving highly non-trivial problems in convex geometry and in empirical processes [23, 26, 32, 33].

Using the bounds on the uniform entropy numbers and theorem 2.3, one can establish the following sample complexity estimates.

Theorem 2.19 *There is an absolute constant C such that for every class $F \subset B\big(L_\infty(\Omega)\big)$ and every $0 < \varepsilon, \delta < 1$,*

$$\sup_\mu Pr\Big\{\sup_{f \in F} |\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_\mu f| \geq \varepsilon\Big\} \leq \delta,$$

provided that

$$n \geq \frac{C}{\varepsilon^2}\Big(\mathrm{fat}_{\varepsilon/8}(F) \cdot \log\Big(\frac{2}{\varepsilon}\Big) + \log\Big(\frac{8}{\delta}\Big)\Big).$$

Unfortunately, the VC dimension and the fat-shattering dimension have become the central issue in machine learning literature. One must remember that the combinatorial parameters were introduced as a way to estimate the uniform entropy numbers. If fact, they seem to be the wrong parameters to measure the complexity of learning problems. Ironically, they have a considerable geometric significance as many results indicate.

To sum-up the results we have presented so far, it is possible to obtain uGC sample complexity estimates via symmetrization, a covering argument and Hoeffding's inequality. The combinatorial parameters are used only to estimate the covering numbers one needs. One point in which a slight improvement can be made, is by replacing Hoeffding's inequality with inequalities of a similar nature, (e.g. Bernstein's inequality or Bennett's inequality [34]) in which additional data on the moments of the random variables is used to obtain tighter deviation bounds. However, this does not resolve the main problem in this line of argumentation - that passing to an ε-cover and applying the union bound is **horribly loose**. To solve this problem one needs a stronger deviation inequality for a supremum over a family of functions and not just a single one. This "functional" inequality is the subject of the next section and we show it yields tighter complexity bounds.

2.3 Talagrand's inequality

Let us begin by recalling Bernstein's inequality [18, 34].

Theorem 2.20 *Let μ be a probability measure on Ω and let $X_1, ..., X_n$ be independent random variables distributed according to μ. Given a function $f : \Omega \to \mathbb{R}$, set $Z = \sum_{i=1}^n f(X_i)$, let $b = \|f\|_\infty$ and put $v = \mathbb{E}(\sum_{i=1}^n f^2(X_i))$. Then,*

$$Pr\big\{|Z - \mathbb{E}_\mu Z| \geq x\big\} \leq 2e^{-\frac{x^2}{2(v+bx/3)}}.$$

This deviation result is tighter than Hoeffding's inequality because one has additional data on the variance of the random variable Z, which leads to potentially sharper bounds. It has been a long standing open question whether a similar result can be obtained when replacing Z by $\sup_{f\in F}|\sum_{i=1}^n\big(f(X_i) - \mathbb{E}_\mu f\big)|$. This "functional" inequality was first established by Talagrand [33], and later was modified and partially improved by Ledoux [15], Massart [18], Rio [28] and Bousquet [3].

Theorem 2.21 *[18] Let μ be a probability measure on Ω and let $X_1, ..., X_n$ be independent random variables distributed according to μ. Given a class of functions F, set $Z = \sup_{f\in F}|\sum_{i=1}^n\big(f(X_i) - E_\mu f\big)|$, let $b = \sup_{f\in F}\|f\|_\infty$ and put $\sigma^2 = \sup_{f\in F}\sum_{i=1}^n \mathrm{var}\big(f(X_i)\big)$. Then, there is an absolute constant $C \geq 1$ such that for every $x > 0$ there is a set of probability larger than $1 - e^{-x}$ on which*

$$Z \leq 2\mathbb{E}Z + C(\sigma\sqrt{x} + bx). \tag{2.6}$$

Observe that if F consists of functions which are bounded by 1 then $b = 1$ and $\sigma \leq \sqrt{n}$. If we select $x = n\varepsilon^2/4C^2$ then with probability larger than $1 - e^{-\frac{n\varepsilon^2}{4C^2}}$,

$$\sup_{f\in F}|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_\mu f| \leq 2\mathbb{E}\sup_{f\in F}|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_\mu f| + \frac{3\varepsilon}{4}.$$

This equation holds with probability larger than $1 - \delta$ provided that $n \geq (4C^2/\varepsilon^2)\log\frac{1}{\delta}$.

It follows that the dominating term in the complexity estimate is the expectation of the random variable Z. Again, the notion of symmetrization will come to our rescue in the attempt to estimate $\mathbb{E}Z$. Let us define the (global) Rademacher averages associated with a class of functions.

Definition 2.22 *Let μ be a probability measure on Ω and set F to be a class of uniformly bounded functions. For every integer n, let*

$$R_n(F) = \mathbb{E}_\mu \mathbb{E}_\varepsilon \frac{1}{\sqrt{n}} \sup_{f \in F} |\sum_{i=1}^n \varepsilon_i f(X_i)|,$$

where $(X_i)_{i=1}^n$ are independent random variables distributed according to μ and $(\varepsilon_i)_{i=1}^n$ are independent Rademacher random variables.

The reason for the seemingly strange normalization (of $1/\sqrt{n}$ instead of $1/n$) will become evident in the next section. Now, we can prove an "averaged" version of the symmetrization result:

Theorem 2.23 *Let μ be a probability measure and set F to be a class of functions on Ω. Denote*

$$Z = \sup_{f \in F} |\frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}_\mu f|,$$

where $(X_i)_{i=1}^n$ are independent random variables distributed according to μ. Then,

$$\mathbb{E}_\mu Z \leq 2 \frac{R_n(F)}{\sqrt{n}} \leq 4\mathbb{E}_\mu Z + 2\big|\sup_{f \in F} \mathbb{E}_\mu f\big| \cdot \mathbb{E}_\varepsilon |\frac{1}{n} \sum_{i=1}^n \varepsilon_i|.$$

Proof: Let $Y_1, ..., Y_n$ be an independent copy of $X_1, ..., X_n$. Then,

$$\mathbb{E}_X \sup_{f \in F} \big|\frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}_Y f\big| = \mathbb{E}_X \sup_{f \in F} \big|\frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}_Y f - \mathbb{E}_Y \big(\frac{1}{n} \sum_{i=1}^n f(Y_i) - \mathbb{E}_Y f\big)\big| = (1)$$

Conditioning (1) with respect to $X_1, ..., X_n$ and then applying Jensen's inequality with respect to $\mathbb{E}_Y$ and Fubini's Theorem, it follows that

$$(1) \leq \frac{1}{n} \mathbb{E}_X \mathbb{E}_Y \sup_{f \in F} \big|\sum_{i=1}^n f(X_i) - \sum_{i=1}^n f(Y_i)\big| = \frac{1}{n} \mathbb{E}_X \mathbb{E}_Y \sup_{f \in F} \big|\sum_{i=1}^n \varepsilon_i \big(f(X_i) - f(Y_i)\big)\big|,$$

where the latter inequality holds for every $(\varepsilon_i)_{i=1}^n \in \{-1, 1\}^n$. Therefore, it also holds when taking the expectation with respect to the Rademacher random variables $(\varepsilon_i)_{i=1}^n$. By the triangle inequality,

$$\frac{1}{n} \mathbb{E}_X \mathbb{E}_Y \mathbb{E}_\varepsilon \sup_{f \in F} \big|\sum_{i=1}^n \varepsilon_i \big(f(X_i) - f(Y_i)\big)\big| \leq \frac{2}{n} \mathbb{E}_X \mathbb{E}_\varepsilon \sup_{f \in F} \big|\sum_{i=1}^n \varepsilon_i f(X_i)\big| = \frac{2R_n(F)}{\sqrt{n}}.$$

To prove the upper bound, the starting point is the triangle inequality which yields that

$$\frac{1}{n}\mathbb{E}_X\mathbb{E}_\varepsilon \sup_{f\in F}\Big|\sum_{i=1}^n \varepsilon_i f(X_i)\Big| \le \frac{1}{n}\mathbb{E}_X\mathbb{E}_\varepsilon \sup_{f\in F}\Big|\sum_{i=1}^n \varepsilon_i\big(f(X_i)-\mathbb{E}_\mu f\big)\Big| + \big|\sup_{f\in F}\mathbb{E}_\mu f\big| \cdot \mathbb{E}_\varepsilon\Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\Big|.$$

To estimate the first term, let (Z_i) be the stochastic process defined by $Z_i(f) = f(X_i) - \mathbb{E}_\mu f$ and let W_i be an independent copy of (Z_i). For every $f \in F$, $\mathbb{E}W_i(f) = 0$, thus,

$$\begin{aligned}\mathbb{E}_X\mathbb{E}_\varepsilon \sup_{f\in F}\Big|\sum_{i=1}^n \varepsilon_i\big(f(X_i)-\mathbb{E}_\mu f\big)\Big| &= \mathbb{E}_Z\mathbb{E}_\varepsilon \sup_{f\in F}\Big|\sum_{i=1}^n \varepsilon_i Z_i(f)\Big| \\ &= \mathbb{E}_\varepsilon\mathbb{E}_Z \sup_{f\in F}\Big|\sum_{i=1}^n \varepsilon_i\big(Z_i(f)-\mathbb{E}_W W_i(f)\big)\Big|.\end{aligned}$$

For every realization of the Rademacher random variables $(\varepsilon_i)_{i=1}^n$ and by Jensen's inequality conditioned with respect to the Z_i,

$$\mathbb{E}_Z \sup_{f\in F}\Big|\sum_{i=1}^n \varepsilon_i\big(Z_i(f)-\mathbb{E}_W W_i(f)\big)\Big| \le \mathbb{E}_Z\mathbb{E}_W \sup_{f\in F}\Big|\sum_{i=1}^n \varepsilon_i\big(Z_i(f)-W_i(f)\big)\Big|,$$

which is invariant for under any selection of signs ε_i. Therefore,

$$\begin{aligned}\mathbb{E}_\varepsilon\mathbb{E}_Z \sup_{f\in F}\Big|\sum_{i=1}^n \varepsilon_i\big(Z_i(f)-\mathbb{E}_W W_i(f)\big)\Big| &\le \mathbb{E}_Z\mathbb{E}_W \sup_{f\in F}\Big|\sum_{i=1}^n \big(Z_i(f)-W_i(f)\big)\Big| \\ &\le 2\mathbb{E}_Z \sup_{f\in F}\Big|\sum_{i=1}^n Z_i(f)\Big|,\end{aligned}$$

as claimed.

■

This result implies that the expectation of the deviation of the empirical means from the actual ones is controlled by $R_n(F)/\sqrt{n}$. Therefore, we can formulate the following

Corollary 2.24 *Let μ be a probability measure on Ω, set $F \subset B\big(L_\infty(\Omega)\big)$ and put $\sigma^2 = \sup_{f\in F}\sum_{i=1}^n \mathrm{var}\big(f(X_i)\big)$, where (X_i) are independent random variables distributed according to μ. Then, there is an absolute constant $C \ge 1$ such that for every $x > 0$, there is a set of probability larger than $1 - e^{-x}$ on which*

$$\sup_{f\in F}\Big|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_\mu f\Big| \le \frac{4R_n(F)}{\sqrt{n}} + \frac{C}{n}(\sigma\sqrt{x} + bx). \tag{2.7}$$

In particular, there is an absolute constant C such that if

$$n \ge \frac{C}{\varepsilon^2}\max\Big\{R_n^2(F), \log\frac{1}{\delta}\Big\},$$

then $Pr\big\{\sup_{f\in F}|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_\mu f| \ge \varepsilon\big\} \le \delta$.

After establishing that the random averages control the GC sample complexity, the natural question is how to estimate them. In particular, it is interesting to try and estimate them using the covering numbers and the combinatorial parameters which where investigated in previous sections.

2.4 Random averages, combinatorial parameters and covering numbers

In this section we present several ways in which one can bound the Rademacher averages associated with a class F. Firstly we present structural results, which enable one to compute the averages of complicated classes using those of simple ones. Next, we give an example of a case in which the averages can be computed directly. Finally, we show how estimates on the empirical entropy of a class can be used to bound the random averages.

2.4.1 Structural results

The following theorem summarizes some of the properties of the Rademacher averages we shall use. The difficulty of the proofs of the different observations varies considerably. Some of the claims are straightforward while others are very deep facts. Most of the results are true when replacing the Rademacher random variables with independent standard gaussian ones (with very similar proofs), but we shall not present the analogous result in the gaussian case.

Theorem 2.25 *Let F and G be classes of real-valued functions on (Ω, μ). Then, for every integer n,*

1. *If $F \subset G$, $R_n(F) \leq R_n(G)$.*
2. *$R_n(F) = R_n(\mathrm{conv}\, F) = R_n(\mathrm{absconv} F)$, where $\mathrm{conv}(F)$ is the convex hull of F and $\mathrm{absconv}(F) = \mathrm{conv}(F \cup -F)$ is the symmetric convex hull of F.*
3. *For every $c \in \mathbb{R}$, $R_n(cF) = |c| R_n(F)$.*
4. *If $\phi : \mathbb{R} \to \mathbb{R}$ is a Lipschitz function with a constant L_ϕ and satisfies that $\phi(0) = 0$, then $R_n(\phi \circ F) \leq 2L_\phi R_n(F)$, where $\phi \circ F = \{\phi\big(f(\cdot)\big) | f \in F\}$.*
5. *For every $1 \leq p < \infty$, there is a constant c_p which depend only on p, such that for every $\{x_1, ..., x_n\} \in \Omega$,*
$$c_p\big(\mathbb{E}_\varepsilon \sup_{f \in F} \big|\sum_{i=1}^n \varepsilon_i f(x_i)\big|^p\big)^{\frac{1}{p}} \leq \mathbb{E}_\varepsilon \sup_{f \in F} \big|\sum_{i=1}^n \varepsilon_i f(x_i)\big| \leq \big(\mathbb{E}_\varepsilon \sup_{f \in F} \big|\sum_{i=1}^n \varepsilon_i f(x_i)\big|^p\big)^{\frac{1}{p}}.$$
6. *For any function $h \in L_2(\mu)$, $R_n(F+h) \leq R_n(F) + (\mathbb{E}_\mu h^2)^{\frac{1}{2}}$, where $F+h = \{f+h | f \in F\}$.*

7. For every $1 < p < \infty$ there is an absolute constant c_p for which

$$c_p\big(\mathbb{E}\sup_{f\in F}\big|\sum_{i=1}^n \varepsilon_i f(X_i)\big|^p\big)^{\frac{1}{p}} \leq \mathbb{E}\sup_{f\in F}\big|\sum_{i=1}^n \varepsilon_i f(X_i)\big| \leq \big(\mathbb{E}\sup_{f\in F}\big|\sum_{i=1}^n \varepsilon_i f(X_i)\big|^p\big)^{\frac{1}{p}},$$

provided that $\sup_{f\in F}\mathbb{E}_\mu f^2 \geq 1/n$.

Proof: Parts 1 and 3 are immediate from the definitions. To see part 2, observe that $R_n(F) \leq R_n\big(\mathrm{conv}(F)\big) \leq R_n\big(\mathrm{absconv}(F)\big)$. To prove the reverse inequality, note that $\mathcal{H} = \mathrm{absconv}(F)$ is symmetric and convex. Hence, for every sample $x_1,...,x_n$ and any realization of $(\varepsilon_i)_{i=1}^n$, $\sup_{h\in\mathcal{H}}|\sum_{i=1}^n \varepsilon_i h(x_i)| = \sup_{h\in\mathcal{H}}\sum_{i=1}^n \varepsilon_i h(x_i)$. Since every $h \in \mathcal{H}$ is given by $\sum_{i=1}^m \lambda_j f_j$ where $\sum_{j=1}^m |\lambda_j| = 1$, then

$$\sum_{i=1}^n \varepsilon_i h(x_i) = \sum_{j=1}^m \lambda_j \sum_{i=1}^n \varepsilon_i f_j(x_i) \leq \sup_{f\in F}\big|\sum_{i=1}^n \varepsilon_i f(x_i)\big|.$$

Hence, the supremum with respect to F and to $\mathcal{H}$ coincide.

Part 4 is called the contraction inequality, and is due to Ledoux and Talagrand [16, Corollary 3.17].

Part 5 is the Kahane-Khintchine inequality [25]. As for part 6, note that for every sample $x_1,...,x_n$,

$$\mathbb{E}_\varepsilon \sup_{f\in F}\big|\sum_{i=1}^n \varepsilon_i\big(f(x_i)+h(x_i)\big)\big| \leq \mathbb{E}_\varepsilon \sup_{f\in F}\big|\sum_{i=1}^n \varepsilon_i f(x_i)\big| + \mathbb{E}_\varepsilon\big|\sum_{i=1}^n \varepsilon_i h(x_i)\big| = (*).$$

By Khintchine's inequality for the second term and the fact that $(\varepsilon_i)_{i=1}^n$ are independent,

$$(*) \leq \mathbb{E}_\varepsilon \sup_{f\in F}\big|\sum_{i=1}^n \varepsilon_i f(x_i)\big| + \big(\sum_{i=1}^n h^2(x_i)\big)^{\frac{1}{2}}.$$

Normalizing by $1/\sqrt{n}$, taking the expectation with respect to μ and by Jensen's inequality,

$$R_n(F+h) \leq R_n(F) + (\mathbb{E}_\mu h^2)^{\frac{1}{2}}.$$

Finally, part 7 follows from a concentration argument which will be presented in appendix A.

■

Remark 2.26 *A significant fact we do not use but feel can not go unmentioned is that the gaussian averages and the Rademacher averages are closely connected. Indeed, one* *can show (see, e.g. [25]) that there are absolute constants c and C which satisfy that for every class F, every integer n and any realization $\{x_1, ..., x_n\}$*

$$c\mathbb{E}_\varepsilon \sup_{f \in F} \Big| \sum_{i=1}^{n} \varepsilon_i f(x_i) \Big| \leq \mathbb{E}_\varepsilon \sup_{f \in F} \Big| \sum_{i=1}^{n} g_i f(x_i) \Big| \leq C\mathbb{E}_\varepsilon \sup_{f \in F} \Big| \sum_{i=1}^{n} \varepsilon_i f(x_i) \Big| \cdot \log n.$$

When one tries to estimate the random averages, the first and most natural approach is to try and compute them directly. There are very few cases in which such an attempt would be successful, and the one we chose to present is the case of kernel classes.

2.4.2 Example: Kernel Classes

Assume that Ω is a compact set and let $K : \Omega \times \Omega \to \mathbb{R}$ to be a positive definite, continuous function. Let μ be a probability measure on Ω, and consider the integral operator $T_K : L_2(\mu) \to L_2(\mu)$ given by $T_K f = \int K(x,y) f(y) d\mu(y)$. By Mercer's Theorem, T_K has a diagonal representation, that is, there exists a complete, orthonormal basis of $L_2(\mu)$, which is denoted by $\big(\phi_n(x)\big)_{n=1}^{\infty}$, and a non-increased sequence of eigenvalues $(\lambda_n)_{i=1}^{\infty}$ which satisfy that for every sequence $(a_n) \in \ell_2$, $T_K(\sum_{n=1}^{\infty} a_n \phi_n) = \sum_{n=1}^{\infty} a_n \lambda_n \phi_n$. Under certain mild assumptions on the measure μ, Mercer's Theorem implies that for every $x, y \in \Omega$,

$$K(x,y) = \sum_{n=1}^{\infty} \lambda_n \phi_n(x) \phi_n(y).$$

Let F_K be the class consisting of all the functions of the form $\sum_{i=1}^{m} a_i K(x_i, \cdot)$ for every $m \in \mathbb{N} \cup \{\infty\}$, every $(x_i)_{i=1}^{m} \subset \Omega$ and every sequence $(a_i)_{i=1}^{m}$ for which $\sum_{i,j=1}^{m} a_i a_j K(x_i, x_j) \leq 1$.

One can show that F_K is the unit ball of a Hilbert space associated with the integral operator, called the reproducing kernel Hilbert space, and we denote it by $\mathcal{H}$. In fact, the unit ball of $\mathcal{H}$ is simply $\sqrt{T_K}\big[B\big(L_2(\mu)\big)\big]$, which is the image of the $L_2(\mu)$ unit ball by the operator which maps ϕ_i to $\sqrt{\lambda_i}\phi_i$. An important property of the inner product in $\mathcal{H}$ is that for every $f \in \mathcal{H}$, $\big\langle f, K(x, \cdot) \big\rangle_{\mathcal{H}} = f(x)$.

An alternative way to define the reproducing kernel Hilbert space is via the feature map. Let $\Phi : \Omega \to \ell_2$ be defined by $\Phi(x) = \big(\sqrt{\lambda_i}\phi_i(x)\big)_{i=1}^{\infty}$. Then,

$$F_K = \big\{ f(\cdot) = \big\langle \beta, \Phi(\cdot) \big\rangle_{\mathcal{H}} \big| \|\beta\|_2 \leq 1 \big\}.$$

Observe that for every $x, y \in \Omega$, $\big\langle \Phi(x), \Phi(y) \big\rangle_{\mathcal{H}} = K(x,y)$.

Let us compute the Rademacher averages of F_K with respect to the probability measure μ.

Theorem 2.27 *Assume that the largest eigenvalue of T_K satisfies that $\lambda_1 \geq 1/n$. Then, for every such integer n,*

$$c\Big(\sum_{i=1}^{\infty} \lambda_i\Big)^{\frac{1}{2}} \leq R_n(F_K) \leq C\Big(\sum_{i=1}^{\infty} \lambda_i\Big)^{\frac{1}{2}},$$

where $(\lambda_i)_{i=1}^{\infty}$ are the eigenvalues of the integral operator T_K arranged in a non increasing order, and C, c are absolute constants.

Remark 2.28 *As the proof we present reveals, the upper bound on $R_n(F_K)$ is true even without the assumption on the largest eigenvalue of T_K.*

Before proving the claim, we require the following lemma:

Lemma 2.29 *Let F_K be the unit ball of the reproducing kernel Hilbert space $\mathcal{H}$ associated with the kernel K. For every sample $s_n = \{x_1, ..., x_n\}$ let $\big(\theta_i(s_n)\big)_{i=1}^{n}$ be the singular values of the operator $T : \mathbb{R}^n \to \mathcal{H}$ defined by $Te_i = K(x_i, \cdot)$. Then,*

$$\mathbb{E}_\varepsilon \frac{1}{\sqrt{n}} \sup_{f \in F_K} \big|\sum_{i=1}^{n} \varepsilon_i f(x_i)\big|^2 = \sum_{i=1}^{n} \theta_i^2.$$

Proof: By the reproducing kernel property,

$$\mathbb{E}_\varepsilon \sup_{f \in F_K} \big|\sum_{i=1}^{n} \varepsilon_i f(x_i)\big|^2 = \mathbb{E}_\varepsilon \sup_{f \in F_K} \big|\sum_{i=1}^{n} \big\langle \varepsilon_i K(x_i, \cdot), f\big\rangle_{\mathcal{H}}\big|^2 = \mathbb{E}_\varepsilon \sup_{f \in F_K} \big|\big\langle \sum_{i=1}^{n} \varepsilon_i Te_i, f\big\rangle_{\mathcal{H}}\big|^2.$$

Since F_K is the unit ball in $\mathcal{H}$ then $\mathbb{E}_\varepsilon \sup_{f \in F_K} |\big\langle \sum_{i=1}^{n} \varepsilon_i Te_i, f\big\rangle_{\mathcal{H}}|^2 = \mathbb{E}_\varepsilon \| \sum_{i=1}^{n} \varepsilon_i Te_i\|_{\mathcal{H}}^2$. Thus,

$$\mathbb{E}_\varepsilon \sup_{f \in F_K} \big|\sum_{i=1}^{n} \varepsilon_i f(x_i)\big|^2 = \mathbb{E}_\varepsilon \big\| \sum_{i=1}^{n} \varepsilon_i Te_i \big\|_{\mathcal{H}}^2 = \sum_{i=1}^{n} \|Te_i\|_{\mathcal{H}}^2 = \sum_{i=1}^{n} \theta_i^2(s_n),$$

proving our claim.

$\blacksquare$

Proof of Theorem 2.27: Firstly, it is easy to see that there exists some $f \in F_K$ for which $\mathbb{E}_\mu f^2 \geq 1/n$. Indeed, $f = \sqrt{T_K}\phi_1 = \sqrt{\lambda_1}\phi_1 \in \mathcal{H}$ satisfies that $\mathbb{E}_\mu f^2 = \lambda_1 \geq 1/n$. Thus, using part 7 of theorem 2.25, $R_n(F)$ is equivalent to $n^{-1/2}\big(\mathbb{E}\sup_{f \in F_K} \big|\sum_{i=1}^{n} \varepsilon_i f(X_i)\big|^2\big)^{1/2}$. Applying the previous lemma and using its notation,

$$\mathbb{E}_\mu\Big(\mathbb{E}_\varepsilon \sup_{f \in F} \big|\sum_{i=1}^{n} \varepsilon_i f(X_i)\big|^2 \big| X_1, ..., X_n\Big) = \mathbb{E}_\mu \sum_{i=1}^{n} \theta_i^2(s_n).$$

By the definition of the operator T, $\big(\theta_i^2(s_n)\big)_{i=1}^n$ are the eigenvalues of T^*T, and it is easy to see that $T^*T = \big(K(x_i,x_j)\big)_{i,j=1}^n$. Therefore,

$$\sum_{i=1}^n \theta_i^2(s_n) = \mathrm{tr}(T^*T) = \sum_{i=1}^n K(x_i,x_i).$$

Hence,

$$\mathbb{E}_\mu\big(\mathbb{E}_\varepsilon \sup_{f\in F} \big|\sum_{i=1}^n \varepsilon_i f(X_i)\big|^2 \big| X_1,...,X_n\big) = \mathbb{E}_\mu\big(\sum_{i=1}^n K(X_i,X_i)\big).$$

To conclude the proof, one has to take the expectation with respect to μ and recall that

$$\mathbb{E}_\mu K(X_i,X_i) = \mathbb{E}_\mu \sum_{j=1}^\infty \lambda_j \phi_j^2(X_i) = \sum_{j=1}^\infty \lambda_j.$$

■

Corollary 2.30 *Let (Ω,μ) be a probability space, set F_K to be the kernel class and put $\mathrm{tr}(K)=\sum_{i=1}^\infty \lambda_i$. Let $T\in B\big(L_\infty(\Omega)\big)$ and denote by $\mathcal{L}$ the loss class associated with F_K and T. Then, there is an absolute constant C such that*

$$Pr\Big\{\sup_{f\in\mathcal{L}} \Big|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_\mu f\Big| \geq \varepsilon\Big\} \leq \delta,$$

provided that $n \geq \frac{C}{\varepsilon^2}\max\{1+\mathrm{tr}(K), \log\frac{1}{\delta}\}$.

Proof: The proof follows immediately from corollary 2.24 and the estimates on the Rademacher averages of F_K. Indeed, by theorem 2.25,

$$\begin{aligned} R_n(\mathcal{L}) &= R_n\big((F_K-T)^2 - (P_{F_K}T-T)^2\big) \leq 4R_n(F_K-T) + \|P_{F_K}T-T\|_\infty^2 \\ &\leq 4\big(R_n(F_K) + C\|T\|_\infty + 1\big) \end{aligned}$$

where C is an absolute constant.

■

2.4.3 Entropy and averages

Unfortunately, in the vast majority of cases, it is next to impossible to compute the random averages directly. Thus, one has to resort to alternative routes to estimate the random averages, especially from above - since this is the direction one needs for sample complexity bounds. We show that it is possible to bound the Rademacher and gaussian averages using the empirical L_2 entropy of the class. This follows from results due to Dudley [6] and Sudakov [31]. Originally, the bounds were established from gaussian processes, and later they were extended to the sub-gaussian setup ([8, 34]), which includes Rademacher processes.

Theorem 2.31 *There are absolute constants C and c for which the following holds. For any integer n, any sample $\{x_1, ..., x_n\}$ and every class F,*

$$c \sup_{\varepsilon>0} \varepsilon \log^{\frac{1}{2}} N\big(\varepsilon, F, L_2(\mu_n)\big) \leq \frac{1}{\sqrt{n}} \mathbb{E}_\varepsilon \sup_{f\in F} \Big|\sum_{i=1}^n \varepsilon_i f(x_i)\Big| \leq C \int_0^\infty \log^{\frac{1}{2}} N\big(\varepsilon, F, L_2(\mu_n)\big) d\varepsilon,$$

where μ_n is the empirical measure supported on the sample.

This result implies that if the class is relatively small, then its Rademacher averages are uniformly bounded.

Corollary 2.32 *There is an absolute constant C such that for every Boolean class F with $\mathrm{VC}(F) = d$ and every integer n, $R_n \leq C\sqrt{d}$.*

Proof: Since F is a Boolean class, all of its members are bounded by 1. Thus, for every $\varepsilon \geq 1$ only a single ball of radius ε is needed to cover F. Using the uniform L_2 entropy bound in theorem 2.15 it follows that for every integer n and every empirical measure μ_n,

$$\log N\big(\varepsilon, F, L_2(\mu_n)\big) \leq Cd \log\big(1/\varepsilon\big),$$

and the claim is evident from theorem 2.31.

■

In a similar way one can show that if $F \subset B\big(L_\infty(\Omega)\big)$ has a polynomial fat-shattering dimension with exponent strictly less than 2, it has uniformly bounded Rademacher averages. This is true because one can obtain a uniform L_2-entropy bound for which the entropy integral converges. It is less obvious what can be done if the entropy integral diverges, in which case theorem 2.31 does not apply.

To handle this case, we present a stronger version of Dudley's entropy bound, which will be formulated for gaussian random variables.

Lemma 2.33 *[19] Let μ_n be an empirical measure supported on $\{x_1, ..., x_n\} \subset \Omega$, put $F \subset B\big(L_\infty(\Omega)\big)$ and set $(\varepsilon_k)_{k=0}^\infty$ to be a monotone sequence decreasing to 0 such that $\varepsilon_0 = 1$. Then, there is an absolute constant C such that for every integer N,*

$$\frac{1}{\sqrt{n}} \mathbb{E} \sup_{f\in F} \Big|\sum_{i=1}^n g_i f(x_i)\Big| \leq C \sum_{k=1}^N \varepsilon_{k-1} \log^{\frac{1}{2}} N\big(\varepsilon_k, F, L_2(\mu_n)\big) + 2\varepsilon_N n^{\frac{1}{2}},$$

where $(g_i)_{i=1}^n$ are standard gaussian random variables. In particular,

$$\frac{1}{\sqrt{n}}\mathbb{E}\sup_{f\in F}\big|\sum_{i=1}^n g_i f(x_i)\big| \leq C\sum_{k=1}^N \varepsilon_{k-1}\mathrm{fat}_{\varepsilon_k/8}^{1/2}(F)\log^{\frac{1}{2}}\big(\frac{2}{\varepsilon_k}\big)+2\varepsilon_N n^{\frac{1}{2}}. \tag{2.8}$$

The latter part of lemma 2.33 follows from its first part and theorem 2.18. Before presenting the proof of lemma 2.33, we require the following lemma, which is based on the classical inequality due to Slepian [27, 8].

Lemma 2.34 *Let $(Z_i)_{i=1}^N$ be gaussian random variables (i.e., $Z_i=\sum_{j=1}^m a_j g_j$ where (g_i) are independent standard gaussian random variables). Then, there is some absolute constant C such that $\mathbb{E}\sup_i Z_i \leq C\sup_{i,j}\|Z_i-Z_j\|_2\log^{\frac{1}{2}}N$.*

Proof of lemma 2.33: We may assume that F is symmetric and contains 0. The proof in the non-symmetric case follows the same path. Let μ_n be an empirical measure supported on $\{x_1,...,x_n\}$. For every $f\in F$, let $Z_f=n^{-1/2}\sum_{i=1}^n g_i f(x_i)$, where $(g_i)_{i=1}^n$ are independent standard gaussian random variables on the probability space (Y,P). Set $\mathcal{Z}_{\mathcal{F}}=\{Z_f|f\in\mathcal{F}\}$ and define $V:L_2(\mu_n)\to L_2(Y,P)$ by $V(f)=Z_f$. Since V is an isometry for which $V(F)=\mathcal{Z}_F$ then

$$N\big(\varepsilon,F,L_2(\mu_n)\big)=N\big(\varepsilon,\mathcal{Z}_F,L_2(P)\big).$$

Let $(\varepsilon_k)_{k=0}^\infty$ be a monotone sequence decreasing to 0 such that $\varepsilon_0=1$ and set $H_k\subset\mathcal{Z}_{\mathcal{F}}$ to be a $2\varepsilon_k$ cover of $\mathcal{Z}_F$. Thus, for every $k\in\mathbb{Z}$ and every $Z_f\in\mathcal{Z}_F$ there is some $Z_f^k\in H_k$ such that $\|Z_f-Z_f^k\|_2\leq 2\varepsilon_k$, and we select $Z_f^0=0$. Writing $Z_f=\sum_{k=1}^N(Z_f^k-Z_f^{k-1})+Z_f-Z_f^N$ it follows that

$$\mathbb{E}\sup_{f\in F}Z_f\leq\sum_{k=1}^N\mathbb{E}\sup_{f\in F}(Z_f^k-Z_f^{k-1})+\mathbb{E}\sup_{f\in\mathcal{F}}(Z_f-Z_f^N).$$

By the definition of Z_f^k and lemma 2.34, there is an absolute constant C for which

$$\begin{aligned}\mathbb{E}\sup_{f\in F}(Z_f^k-Z_f^{k-1})\leq&\mathbb{E}\sup\big\{Z_i-Z_j|Z_i\in H_k,Z_j\in H_{k-1},\|Z_i-Z_j\|_2\leq 4\varepsilon_{k-1}\big\}\\ \leq&C\sup_{i,j}\|Z_i-Z_j\|_2\log^{\frac{1}{2}}|H_k|\,|H_{k-1}|\\ \leq&C\varepsilon_{k-1}\log^{\frac{1}{2}}N\big(\varepsilon_k,F,L_2(\mu_n)\big)\ .\end{aligned}$$

Since $Z_f^N\in\mathcal{Z}_F$, there is some $f'\in F$ such that $Z_f^N=Z_{f'}$. Hence,

$$\left(\sum_{i=1}^n\Big(\frac{f(x_i)-f'(x_i)}{\sqrt{n}}\Big)^2\right)^{\frac{1}{2}}=\big\|Z_f-Z_{f'}\big\|_2\leq 2\varepsilon_N,$$

which implies that for every $f \in F$ and every $y \in Y$,

$$\left|Z_f(y) - Z_f^N(y)\right| \leq \sum_{i=1}^n \left|\frac{f(x_i) - f'(x_i)}{\sqrt{n}} g_i(y)\right| \leq 2\varepsilon_N \Big(\sum_{i=1}^n g_i^2(y)\Big)^{\frac{1}{2}} \ .$$

Therefore, $\mathbb{E} \sup_{f \in F} Z_f - Z_f^N \leq \varepsilon_N \mathbb{E}\big(\sum_{i=1}^n g_i^2\big)^{\frac{1}{2}} = 2\varepsilon_N \sqrt{n}$, and the claim follows.

■

Using this result it is possible to estimate the Rademacher averages of classes with a polynomial fat-shattering dimension.

Theorem 2.35 *Let $F \subset B\big(L_\infty(\Omega)\big)$ and assume that there is some $\gamma > 1$ such that for any $\varepsilon > 0$, $\mathrm{fat}_\varepsilon(F) \leq \gamma \varepsilon^{-p}$. Then, there are absolute constants C_p, which depends only on p, such that*

$$R_n(F) \leq C_p \gamma^{\frac{1}{2}} \begin{cases} 1 & \textit{if } 0 < p < 2 \\ \log^{3/2} n & \textit{if } p = 2 \\ n^{\frac{1}{2} - \frac{1}{p}} & \textit{if } p > 2. \end{cases}$$

Proof: Let μ_n be an empirical measure on Ω. If $p < 2$ then by theorem 2.18,

$$\int_0^\infty \log^{\frac{1}{2}} N\big(\varepsilon, F, L_2(\mu_n)\big) d\varepsilon \leq C_p \gamma^{\frac{1}{2}}$$

and the bound follows from the upper bound in theorem 2.31.

Assume that $p \geq 2$ and, using the notation of lemma 2.33, select $\varepsilon_k = 2^{-k}$ and $N = p^{-1} \log n$. By (2.8),

$$\begin{aligned} R_n(F) &\leq C_p \gamma^{\frac{1}{2}} \sum_{i=1}^N \varepsilon_k^{1 - \frac{p}{2}} \log^{\frac{1}{2}} \frac{2}{\varepsilon_k} + 2\varepsilon_N n^{\frac{1}{2}} \\ &\leq C_p \gamma^{\frac{1}{2}} \sum_{i=1}^N \sqrt{k} 2^{k(\frac{p}{2} - 1)} + 2n^{\frac{1}{2} - \frac{1}{p}}. \end{aligned}$$

If $p = 2$, the geometric sum is bounded by

$$C_p \gamma^{\frac{1}{2}} N^2 \leq C_p \gamma^{\frac{1}{2}} \log^{3/2} n,$$

whereas is $p > 2$ it is bounded by $C_p \gamma^{\frac{1}{2}} n^{\frac{1}{2} - \frac{1}{p}}$.

■

These bounds on R_n are "worst case" bounds, since they hold for any empirical measure. In fact, the underlying measure μ plays no part in the bounds. Using a geometric interpretation of the fat-shattering dimension, it is possible to show that the "worst case" bounds we established are tight, in the sense that if $\mathrm{fat}_\varepsilon(F) = \Omega(\varepsilon^{-p})$ for $p > 2$, then for every integer n there will be a sample $\{x_1, ..., x_n\}$ for which

$$\frac{1}{\sqrt{n}} \mathbb{E}_\varepsilon \sup_{f \in F} \Big| \sum_{i=1}^{n} \varepsilon_i f(x_i) \Big| \geq c n^{\frac{1}{2} - \frac{1}{p}},$$

where c is an absolute constant. Since this is not the main issue we wish to address in these notes, we refer the interested reader to [19].

The complexity bounds that one obtains using corollary 2.24 and theorem 2.35 are a significant improvement to the ones obtained via theorem 2.19. Indeed, the sample complexity estimate obtained there was that if $\mathrm{fat}_\varepsilon(F) = O(\varepsilon^{-p})$ then

$$S_F(\varepsilon, \delta) = O\big(\frac{1}{\varepsilon^{2+p}} \cdot (\log^2 \frac{2}{\varepsilon} + \log \frac{2}{\delta})\big).$$

Using Talagrand's inequality, we obtain a sharper bound:

Theorem 2.36 *Let $F \subset B\big(L_\infty(\Omega)\big)$ and assume that $\mathrm{fat}_\varepsilon(F) \leq \gamma \varepsilon^{-p}$. Then, there is a constant C_p, which depends only on p, such that*

$$S_F(\varepsilon, \delta) \leq C_p \max\big\{\frac{1}{\varepsilon^p}, \frac{1}{\varepsilon^2} \log \frac{1}{\delta}\big\}$$

if $p \neq 2$. If $p = 2$ there is an additional logarithmic factor in $\frac{1}{\varepsilon}$.

We were able to obtain this improved result is because we removed the major looseness-the union bound in the "classical" argument. But this is not the end of the story.... There is still one additional source of sub-optimality; as we said in the introduction, using the uGC property only yields upper bounds to the quantity we wish to explore - the learning sample complexity. In the next section, we use very similar methods to the ones used here and obtain even tighter bounds.

3 Learning sample complexity

After bounding the uGC sample complexity using corollary 2.24 and establishing bounds on the Rademacher averages, we now turn to the alternative approach which will prove to yield tighter learning sample complexity bounds.

Recall that the question we wish to answer is how to ensure that an "almost minimizer" of the empirical loss will be close to the minimum of the actual loss.

Thus, our aim is to bound

$$Pr\big\{\exists f \in \mathcal{L},\ \frac{1}{n} \sum_{i=1}^{n} f(X_i) \leq \varepsilon/2, \mathbb{E}_\mu f \geq \varepsilon\big\}. \tag{3.1}$$

To that end, we need to impose an important structural assumption on the class at hand.

Assumption 3.1 *Assume that there is an absolute constant B such that for every $f \in F$, $\mathbb{E}_\mu f^2 \leq B\mathbb{E}_\mu f$.*

Though this assumption seems restrictive, it turns out that it holds in all the cases we are interested in.

Lemma 3.1 *Let $F \subset B\big(L_\infty(\Omega)\big)$ which satisfies assumption 3.1. Fix $\varepsilon > 0$ and define*

$$H = \big\{\frac{\varepsilon f}{\mathbb{E}_\mu f} | f \in F,\ \mathbb{E}_\mu f \geq \varepsilon,\ \mathbb{E}_\mu f^2 \geq \varepsilon\big\}, \tag{3.2}$$

and set

$$F_\varepsilon = \big\{f \in F|\ \mathbb{E}_\mu f^2 \leq \varepsilon\big\}, \qquad H_\varepsilon = \big\{h \in H|\ \mathbb{E}_\mu h^2 \leq B\varepsilon\big\}.$$

Then,

$$\begin{aligned}&Pr\big\{\exists f \in F,\ \frac{1}{n}\sum_{i=1}^n f(X_i) \leq \varepsilon/2,\ \mathbb{E}_\mu f \geq \varepsilon\big\} \leq \\ &Pr\big\{\sup_{f \in F_\varepsilon} |\mathbb{E}_\mu f - \mathbb{E}_{\mu_n} f| \geq \frac{\varepsilon}{2}\big\} + Pr\Big\{\sup_{h \in H_\varepsilon} |\mathbb{E}_\mu h - \mathbb{E}_{\mu_n} h| \geq \frac{\varepsilon}{2}\Big\}\end{aligned}$$

In particular, for every $0 < \delta < 1$,

$$C_{\mathcal{L}}(\frac{\varepsilon}{2}, \delta) \leq \max\big\{S_{F_\varepsilon}\big(\frac{\varepsilon}{2}, \frac{\delta}{2}\big), S_{H_\varepsilon}\big(\frac{\varepsilon}{2}, \frac{\delta}{2}\big)\big\}.$$

Proof: Denote by μ_n the random empirical measure $n^{-1}\sum_{i=1}^n \delta_{X_i}$. Then,

$$\begin{aligned}&Pr\big\{\exists f \in F,\ \mathbb{E}_{\mu_n} f \leq \varepsilon/2,\ \mathbb{E}_\mu f \geq \varepsilon\big\} \leq \\ &Pr\big\{\exists f \in F,\ \mathbb{E}_\mu f \geq \varepsilon,\ \mathbb{E}_\mu f^2 < \varepsilon,\ \mathbb{E}_{\mu_n} f \leq \varepsilon/2\big\} + \\ &Pr\big\{\exists f \in F,\ \mathbb{E}_\mu f \geq \varepsilon,\ \mathbb{E}_\mu f^2 \geq \varepsilon,\ \mathbb{E}_{\mu_n} f \leq \varepsilon/2\big\} \\ &= (1) + (2).\end{aligned}$$

If $\mathbb{E}_\mu f \geq \varepsilon$ then $\mathbb{E}_\mu f \geq \frac{1}{2}(\mathbb{E}_\mu f + \varepsilon) \geq \frac{1}{2}\mathbb{E}_\mu f + \mathbb{E}_{\mu_n} f$. Therefore, $|\mathbb{E}_\mu f - \mathbb{E}_{\mu_n} f| \geq \frac{1}{2}\mathbb{E}_\mu f \geq \varepsilon/2$, hence,

$$\begin{aligned}(1) + (2) \leq\ & Pr\big\{\exists f \in F,\ \mathbb{E}_\mu f^2 < \varepsilon,\ |\mathbb{E}_\mu f - \mathbb{E}_{\mu_n} f| \geq \frac{\varepsilon}{2}\big\} \\ &+ Pr\big\{\exists f \in F,\ \mathbb{E}_\mu f \geq \varepsilon,\ \mathbb{E}_\mu f^2 \geq \varepsilon,\ |\mathbb{E}_\mu f - \mathbb{E}_{\mu_n} f| \geq \frac{1}{2}\mathbb{E}_\mu f\big\} = (3) + (4).\end{aligned}$$

The first term is bounded by $Pr\big\{\sup_{f \in F_\varepsilon} |\mathbb{E}_{\mu_n} f - \mathbb{E}_\mu f| \geq \varepsilon/2\big\}$. As for the second, assume that $|\mathbb{E}_{\mu_n} f - \mathbb{E}_\mu f| \geq (\mathbb{E}_\mu f)/2$ and that $\mathbb{E}_\mu f \geq \varepsilon$. Then, $h = \varepsilon f / E_\mu f$ satisfies that $|\mathbb{E}_{\mu_n} h - \mathbb{E}_\mu h| \geq \varepsilon/2$ and since $\mathbb{E}_\mu f^2 \leq B(\mathbb{E}_\mu f)$ then

$$\mathbb{E}_\mu h^2 \leq B\frac{\varepsilon^2}{\mathbb{E}_\mu f} \leq B\varepsilon.$$

Therefore, (4) $\leq Pr\{\exists h \in H_\varepsilon,\ |\mathbb{E}_{\mu_n} h - \mathbb{E}_\mu h| \geq \varepsilon/2\}$.

■

To simplify this estimate, we require the following definition:

Definition 3.2 *Let X be a normed space and let $A \subset X$. We say that A is star-shaped with center x if for every $a \in A$ the interval $[a, x] = \{tx + (1-t)a | 0 \leq t \leq 1\} \subset A$. Given A and x, denote by* star(A, x) *the union of all the intervals $[a, x]$, where $a \in A$.*

It is easy to see that each element $h \in H$ is given by $\alpha_f f$, where $0 \leq \alpha_f \leq 1$. Thus, $H \subset \mathrm{star}(F, 0)$ and obviously $F \subset \mathrm{star}(F, 0)$. Therefore,

$$\begin{aligned} & Pr\{\exists f \in F,\ \frac{1}{n}\sum_{i=1}^n f(X_i) < \varepsilon/2, \mathbb{E}_\mu f \geq \varepsilon\} \leq \\ & 2Pr\Big\{\exists h \in \mathrm{star}(F, 0),\ \mathbb{E}_\mu h^2 \leq B\varepsilon,\ |\mathbb{E}_\mu h - \mathbb{E}_{\mu_n} h| \geq \frac{\varepsilon}{2}\}. \end{aligned} \tag{3.3}$$

This implies that the question of obtaining sample complexity estimates may be reduced to a GC deviation problem for a class which is the intersection of star$(F, 0)$ with an $L_2(\mu)$ ball, centered at 0 with radius proportional to the square-root of the required deviation. Combining this with corollary 2.24 yields the following fundamental result:

Theorem 3.3 *Let $F \subset B\big(L_\infty(\Omega)\big)$ and assume that assumption 3.1 holds. Set $\mathcal{H} =$* star$(F, 0)$ *and for every $\varepsilon > 0$ let $\mathcal{H}_\varepsilon = \mathcal{H} \cap \{h : \mathbb{E}_\mu h^2 \leq \varepsilon\}$. Then, for every $0 < \varepsilon, \delta < 1$,*

$$Pr\Big\{\exists f \in F,\ \frac{1}{n}\sum_{i=1}^n f(X_i) \leq \varepsilon/2,\ \mathbb{E}_\mu f \geq \varepsilon\Big\} \leq \delta$$

Provided that

$$n \geq C \max\{\frac{R_n^2(\mathcal{H}_\varepsilon)}{\varepsilon^2}, \frac{B \log \frac{2}{\delta}}{\varepsilon}\}.$$

The proof of this theorem follows immediately from theorem 2.21.

Theorem 3.3 shows that the important quantity which governs the learning sample complexity is the "localized" Rademacher average $R_n(\mathcal{H}_\varepsilon)$, assuming, of course, that assumption 3.1 holds.

Before presenting bounds on the localized Rademacher averages of some classes, let us comment on assumption 3.1. Assumption 3.1 clearly holds for 2-loss classes if the target function is a member of the original class G, since in that case, $P_G T = T$, and every loss function is nonnegative and bounded by 4. The situation when $T \notin G$ is much more difficult. One can show that if $G \subset B\big(L_\infty(\Omega)\big)$ is convex and $T \in B\big(L_\infty(\Omega)\big)$, then for every probability measure μ and every 2-loss function f, $\mathbb{E}_\mu f^2 \leq 16\mathbb{E}_\mu f$ [17, 20]. In fact, it is possible to obtain results of a similar flavor for q-loss classes, where the "usual" exponent

2 is replaced with some $q \geq 2$ (see [20]). Even the convexity assumption can be relaxed in the following sense; if $G \subset L_2(\mu)$ is not convex, then there will be functions which have more than a single best approximation in G. The set of functions which do not have a unique best approximation in G is denoted by $\mathrm{nup}(G,\mu)$ and it clearly depends on the probability measure μ, because a change of measure generates a different way of measuring distances. One can show ([24]) that given a measure μ and a target $T \notin \mathrm{nup}(G,\mu)$, the 2-loss class $\mathcal{L}$ satisfies that $\mathbb{E}_\mu f^2 \leq B\mathbb{E}_\mu f$ for every $f \in \mathcal{L}$. The constant B will depend on "how far" T is from $\mathrm{nup}(G,\mu)$. Thus, the complexity bounds one obtains in this case are both target and measure dependent.

For the sake of simplicity, in all the cases we shall be interested in we impose the assumption that either $T \in G$, or that G is convex. In both these cases, a selection of $B = 16$ suffices to ensure that assumption 3.1 holds.

3.1 Localized random averages

In an analogous way to what we did in section 2.4, we present two paths one can take when computing the random averages. For the direct approach we present the example of kernel classes. The second approach, which may be used in the vast majority of examples is to apply uniform entropy estimates.

3.1.1 Localized averages of kernel classes

Here, we present a direct tight bound on the localized Rademacher averages of F_K in terms of the eigenvalues of the integral operator T_K. It is important to note that the underlying measure in the definition of R_n and of T_K has to be the same, which emphasizes the difficulty from the learning theoretic viewpoint, since one does not have *a priori* knowledge on the underlying measure.

Theorem 3.4 *[21] There are absolute constants c and C for which the following holds. Let K be a kernel and set μ to be a probability measure on Ω. If $(\lambda_i)_{i=1}^\infty$ are the eigenvalues of the integral operator T_K (with respect to μ) and if $\lambda_1 \geq 1/n$, then for every $\varepsilon \geq 1/n$,*

$$c\Big(\sum_{j=1}^{\infty} \min\{\lambda_i, \varepsilon\}\Big)^{\frac{1}{2}} \leq \frac{1}{\sqrt{n}}\mathbb{E}_\mu\mathbb{E}_\varepsilon \sup_{f\in F_\varepsilon} \Big|\sum_{i=1}^{n} \varepsilon_i f(X_i)\Big| \leq C\Big(\sum_{j=1}^{\infty} \min\{\lambda_i, \varepsilon\}\Big)^{\frac{1}{2}},$$

where $F_\varepsilon = \{f \in F_K,\ \mathbb{E}_\mu f^2 \leq \varepsilon\}$

Remark 3.5 *The upper bound in theorem 3.4 holds even without the assumptions on λ_1 and ε, and this is the direction we require for sample complexity bounds. The assumption is imposed only to enable one to obtain matching upper and lower bounds.*

Proof: Let $R_\varepsilon = \sup_{f \in F_\varepsilon} \left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|$. Just as in the proof of theorem 2.27, there is some $f \in F_K$ for which $\mathbb{E}_\mu f^2 \geq 1/n$. Hence, there will be some $0 < t \leq 1$ for which $f_1 = tf \in F_\varepsilon$ and $\mathbb{E}_\mu f_1^2 \geq 1/n$. Thus, $\sup_{f \in F_\varepsilon} \mathbb{E}_\mu f^2 \geq 1/n$ and by theorem 2.25, part 7, $\mathbb{E}R_\varepsilon$ is equivalent to $(\mathbb{E}R_\varepsilon^2)^{1/2}$.

We can assume that ℓ_2 is the reproducing kernel Hilbert space and recall that $F_K = \{f(\cdot) = \langle \beta, \Phi(\cdot)\rangle | \, \|\beta\|_2 \leq 1\}$, where Φ is the kernel feature map. By Setting $B(\varepsilon) = \{f | \mathbb{E}_\mu f^2 \leq \varepsilon\}$ it follows that $f \in F_K$ is also in $B(\varepsilon)$ if and only if its representing vector β satisfies that $\sum_{i=1}^\infty \beta_i^2 \lambda_i \leq \varepsilon$. Hence, in ℓ_2,

$$F_\varepsilon = F_K \cap B(\varepsilon) = \{\beta | \sum_{i=1}^\infty \beta_i^2 \leq 1, \ \sum_{i=1}^\infty \beta_i^2 \lambda_i \leq \varepsilon\}.$$

Let $\mathcal{E} \subset \ell_2$ be defined as $\{\beta | \sum_{i=1}^\infty \mu_i \beta_i^2 \leq 1\}$, where $\mu_i = (\min\{1, \varepsilon/\lambda_i\})^{-1}$ and note that

$$\mathcal{E} \subset F_K \cap B(\varepsilon) \subset \sqrt{2}\mathcal{E}.$$

Therefore, one can replace F_ε by $\mathcal{E}$ in the computation of $R_n(F_\varepsilon)$, losing a factor of $\sqrt{2}$ at the most. Finally,

$$\mathbb{E} \sup_{\beta \in \mathcal{E}} |\langle \beta, \sum_{j=1}^n \varepsilon_j \Phi(X_j)\rangle|^2 = \mathbb{E} \sup_{\beta \in \mathcal{E}} |\langle \sum_{i=1}^\infty \sqrt{\mu_i}\beta_i e_i, \sum_{i=1}^\infty \big(\frac{\lambda_i}{\mu_i}\big)^{\frac{1}{2}} \big(\sum_{j=1}^n \varepsilon_j \phi_i(X_j)\big) e_i\rangle|^2$$
$$= \mathbb{E}\big(\sum_{i=1}^\infty \frac{\lambda_i}{\mu_i} \sum_{j=1}^\infty \varepsilon_j \phi_i(X_j)\big)^2 = \mathbb{E}_\mu \sum_{i,j} \frac{\lambda_i}{\mu_i} \phi_i^2(X_j) = n \sum_{i=1}^\infty \frac{\lambda_i}{\mu_i},$$

which proves our claim.

■

As an example, consider the case where the eigenvalues of T_K are $\lambda_i \sim 1/i^p$, for some $p > 1$. It is easy to see that in that case, $R_n(F_\varepsilon) \leq C\varepsilon^{1/2 - 1/p}$. Therefore, if $T \in F_K$, then according to theorem 3.3 the learning sample complexity (when the sampling is done with respect to the measure μ!!!) is

$$C(\varepsilon, \delta) = O\Big(\max\Big\{\frac{1}{\varepsilon^{1+1/p}}, \frac{\log(2/\delta)}{\varepsilon}\Big\}\Big).$$

3.1.2 Using the Entropy

The previous section is somewhat misleading since the reader might develop the feeling that computing localized averages directly is a winning strategy. Unfortunately, even if the geometry of the original class is well behaved and enables direct computation, the problem becomes considerably harder in the localized case. In the latter, one has to take into account the intersection body of the original class and an $L_2(\mu)$ ball. Thus, in most cases one has no choice but to resort to indirect methods, like entropy based bounds.

Theorem 2.31 may be used to compute the localized version of the Rademacher averages in the following manner; let Y be a random variable which measures the empirical radius of the class, which is $(\sup_{f\in F} n^{-1}\sum_{i=1}^n f^2(X_i))^{1/2}$. Given a sample $\{x_1,...,x_n\}$ and any $\varepsilon \geq Y^{1/2}(x_1,...,x_n)$, only a single ball is needed to cover the entire class. Hence,

$$\frac{1}{\sqrt{n}}\sup_{f\in F}\Big|\sum_{i=1}^n \varepsilon_i f(x_i)\Big| \leq C\int_0^{Y^{1/2}(x_1,...,x_n)} \log^{\frac{1}{2}} N\big(\varepsilon,F,L_2(\mu_n)\big)d\varepsilon.$$

Taking the expectation with respect to the sample it follows that there is an absolute constant C such that for every class F,

$$R_n(F)\leq C\mathbb{E}\int_0^{\sqrt{Y}} \log^{\frac{1}{2}} N\big(\varepsilon,F,L_2(\mu_n)\big)d\varepsilon.$$

where $Y=\sup_{f\in F} n^{-1}\sum_{i=1}^n f^2(X_i)$.

Of course, the information we have is not on the random variable Y, but rather on $\sigma_F^2=\sup_{f\in F}\mathbb{E}_\mu f^2$. Fortunately, it is possible to connect the two, as the following result which is due to Talagrand [33], shows.

Lemma 3.6 *Let $F\subset B\big(L_\infty(\Omega)\big)$ and set $\sigma_F^2=\sup_{f\in F}\mathbb{E}_\mu f^2$. Then,*

$$\mathbb{E}_\mu\sup_{f\in F}\sum_{i=1}^n f^2(X_i)\leq n\sigma_F^2+8\sqrt{n}R_n(F)$$

Using this fact, it turn out that if one has data on the uniform entropy, one can estimate the localized Rademacher averages. As an example, consider the case when the entropy is logarithmic in $1/\varepsilon$.

Lemma 3.7 *Let $F\subset B\big(L_\infty(\Omega)\big)$ and set $\sigma_F^2=\sup_{f\in F}\mathbb{E}_\mu f^2$. Assume that there are $\gamma>1$, $d\geq 1$ and $p\geq 1$ such that*

$$\log N_2(\varepsilon,F)\leq d\log^p\Big(\frac{\gamma}{\varepsilon}\Big).$$

Then, there is a constant $C_{p,\gamma}$ which depend only on p and γ for which

$$R_n(F)\leq C_{p,\gamma}\max\Big\{\frac{d}{\sqrt{n}}\log^p\frac{1}{\sigma_F},\sqrt{d}\tau\log^{\frac{p}{2}}\frac{1}{\sigma_F}\Big\}.$$

Before proving the lemma, we require the next result:

Lemma 3.8 *For every $0 \le p < \infty$ and $\gamma > 1$, there is some constant $c_{p,\gamma}$ such that for every $0 < x < 1$,*

$$\int_0^x \log^p \frac{\gamma}{\varepsilon} d\varepsilon \le 2x \log^p \frac{c_{p,\gamma}}{x},$$

and $x^{1/2} \log^p \frac{c_{p,\gamma}}{x}$ is increasing and concave in $(0,10)$.

The first part of the proof follows from the fact that both terms are equal at $x = 0$, but for an appropriate constant $c_{p,\gamma}$, the derivative of the function on left-hand side is smaller than that of the function on the right-hand one. The second part is evident by differentiation.

Proof of lemma 3.7: Set $Y = n^{-1} \sup_{f \in F} \sum_{i=1}^n f^2(X_i)$. By theorem 2.31 there is an absolute constant C such that

$$\frac{1}{\sqrt{n}} \mathbb{E}_\varepsilon \sup_{f \in F} \Big| \sum_{i=1}^n \varepsilon_i f(X_i) \Big| \le C \int_0^{\sqrt{Y}} \log^{\frac{1}{2}} N\big(\varepsilon, F, L_2(\mu_n)\big) d\varepsilon = C\sqrt{d} \int_0^{\sqrt{Y}} \log^{\frac{p}{2}} \frac{\gamma}{\varepsilon} d\varepsilon.$$

By lemma 3.8 there is a constant $c_{p,\gamma}$ such that for every $0 < x \le 1$,

$$\int_0^x \log^{\frac{p}{2}} \frac{\gamma}{\varepsilon} d\varepsilon \le 2x \log^{\frac{p}{2}} \frac{c_{p,\gamma}}{x},$$

and $v(x) = \sqrt{x} \log^{p/2}(c_{p,\gamma}/x)$ is increasing and concave in $(0,10)$.

Since $Y \le 1$,

$$\frac{1}{\sqrt{n}} \mathbb{E}_\varepsilon \sup_{f \in F} \Big| \sum_{i=1}^n \varepsilon_i f(X_i) \Big| \le C_p \sqrt{dY} \log^{\frac{p}{2}} \frac{c_{p,\gamma}}{Y},$$

and since $\sigma_F^2 + 8R_n/\sqrt{n} \le 9$, then by Jensen's inequality, lemma 3.6 and the fact that v is increasing in $(0,10)$,

$$\begin{aligned}
\mathbb{E}_\mu \big(Y^{\frac{1}{2}} \log^{\frac{p}{2}} \frac{c_{p,\gamma}}{Y} \big) &\le (\mathbb{E}_\mu Y)^{\frac{1}{2}} \log^{\frac{p}{2}} \frac{c_{p,\gamma}}{\mathbb{E}_\mu Y} \\
&\le c_{p,\gamma} \big(\sigma_F^2 + 8\frac{R_n}{\sqrt{n}} \big)^{\frac{1}{2}} \log^{\frac{p}{2}} \frac{1}{\sigma_F^2 + \frac{8R_n}{\sqrt{n}}} \\
&\le c_{p,\gamma} \big(\sigma_F^2 + \frac{8R_n}{\sqrt{n}} \big)^{\frac{1}{2}} \log^{\frac{p}{2}} \frac{1}{\sigma_F}.
\end{aligned}$$

Therefore,

$$R_n(F) \le C_{p,\gamma} \sqrt{d} \big(\sigma_F^2 + \frac{R_n}{\sqrt{n}} \big)^{\frac{1}{2}} \log^{\frac{p}{2}} \frac{1}{\sigma_F},$$

and our claim follows from a straightforward computation.

■

In a similar manner one can show that if that there are γ and $p < 2$ such that

$$\log N_2(\varepsilon, F) \leq \frac{\gamma}{\varepsilon^p}$$

then

$$R_n(F) \leq C_{p,\gamma} \max\left\{n^{-\frac{1}{2}\frac{2-p}{2+p}}, \sigma_F^{1-\frac{p}{2}}\right\}, \tag{3.4}$$

and if

$$\log N_2(\varepsilon, F) \leq \frac{\gamma}{\varepsilon^p} \log^2 \frac{2}{\varepsilon}$$

then

$$R_n(F) \leq C_{p,\gamma} \max\left\{n^{-\frac{1}{2}\frac{2-p}{2+p}} \log^\beta \frac{2}{\sigma_F}, \sigma_F^{1-\frac{p}{2}} \log \frac{2}{\sigma_F}\right\}, \tag{3.5}$$

where $\beta = 4/(2+p)$.

Let $F \subset B\bigl(L_\infty(\Omega)\bigr)$ and set $F_\varepsilon = \bigl\{f \in F \big| \mathbb{E}_\mu f^2 \leq \varepsilon\bigr\}$. Since $F_\varepsilon \subset F$ then its entropy must be smaller than that of F. Therefore, all the estimates above hold for F_ε when one replaces σ_F^2 by ε.

The next step is to connect the entropy of the original class G to that of $F = \mathrm{star}(\mathcal{L}, 0)$. Let us recall that the uniform entropy for the loss class is controlled by that of G (see lemma 2.8). Hence, all that remains is to see whether taking the star-shaped hull of $\mathcal{L}$ with 0 increases the entropy by much.

Lemma 3.9 *Let X be a normed space and let $A \subset B(X)$ be totally bounded. Then, for any $\|x\| \leq 1$ and every $\varepsilon > 0$,*

$$\log N\bigl(2\varepsilon, \mathrm{star}(A, x)\bigr) \leq \log \frac{2}{\varepsilon} + \log N\bigl(\varepsilon, A\bigr).$$

Proof: Fix $\varepsilon > 0$ and let $y_1, ..., y_k$ be an ε-cover of A. Note that for any $a \in A$ and any $z \in [a, x]$ there is some $z' \in [y_i, x]$ such that $\|z' - z\| < \varepsilon$. Hence, an ε-cover of the union $\cup_{i=1}^n [y_i, z]$ is a 2ε-cover for $\mathrm{star}(A, x)$. Since for every i, $\|x - y_i\| \leq 2$, then each interval may be covered by $2\varepsilon^{-1}$ balls of radius ε and our claim follows.

$\blacksquare$

Corollary 3.10 *Assume that G consists of functions which map Ω into $[0, 1]$ and that the same holds for T. Then, for any $\varepsilon, \rho > 0$,*

$$\log N_2(\rho, F_\varepsilon) \leq \log N_2(\rho/8, G) + \log\bigl(2/\rho\bigr),$$

where $F_\varepsilon = \bigl\{f \in \mathrm{star}(\mathcal{L}, 0) \big| \mathbb{E}_\mu f^2 \leq \varepsilon\bigr\}$.

This result yields sample complexity estimates when one has estimates on the L_2 entropy of the class (which can be obtained using the combinatorial parameters or other methods). The case we present here is when the class has a polynomial uniform entropy.

Theorem 3.11 *Let $G \subset B\big(L_\infty(\Omega)\big)$ be a convex class of functions and assume that $N_2(\varepsilon, G) \leq \gamma\varepsilon^{-p}$ for some $0 < p < \infty$. Set $T \in B\big(L_\infty(\Omega)\big)$ and put $\mathcal{L}$ to be the loss class associated with G and T. Then,*

$$Pr\Big\{\exists f \in \mathcal{L},\ \frac{1}{n}\sum_{i=1}^n f(X_i) \leq \varepsilon, \mathbb{E}_\mu f \geq 2\varepsilon\Big\} \leq \delta,$$

provided that

$$n \geq C(p,\gamma)\max\Big\{\big(\frac{1}{\varepsilon}\big)^{1+\frac{p}{2}}, \frac{\log(1/\delta)}{\varepsilon}\Big\} \quad \text{if } 0 < p < 2,$$

and

$$n \geq C(p,\gamma)\max\Big\{\big(\frac{1}{\varepsilon}\big)^{p}, \frac{\log(1/\delta)}{\varepsilon}\Big\} \quad \text{if } p > 2.$$

Proof: Let $F = \text{star}(\mathcal{L}, 0)$ and set $F_\varepsilon = \big\{f \in F \big| \mathbb{E}_\mu f^2 \leq \varepsilon\big\}$. Applying theorem 2.35 it follows that for every integer n, every $\varepsilon > 0$ and any $p > 2$,

$$R_n(F_\varepsilon) \leq R_n(F) \leq C_p n^{\frac{1}{2}-\frac{1}{p}}.$$

To estimate the localized averages for $0 < p < 2$, one uses the previous corollary and (3.4). Both parts of the theorem are now immediate from theorem 3.3.

■

3.2 The iterative scheme

The biggest downside in the our analysis is the fact that the localized Rademacher averages are very hard to compute, and it almost impossible to estimate them using the empirical data one receives. If fact, all the results presented here were based on some kind of an a-priori data on the learning problem we had to face; for example, we imposed assumptions on the growth rates of the uniform entropy of the class.

It is highly desirable to obtain estimates which are data-dependent. This could be done if we had the ability to replace the $L_2(\mu)$ ball in the definition of the localized averages by the empirical ball $\big\{f \in F \big| n^{-1}\sum_{i=1}^n f(X_i) \leq \varepsilon\big\}$

Koltchinskii and Panchenko [13] have introduced a computable iterative scheme which enabled them to replace the "actual" ball by an empirical one for a random sequence of radii $r_k = r_k(X_1, ..., X_n)$. In some cases, this method proved to be an effective way of bounding the localized averages. In fact, when one has some "global" data (e.g. uniform entropy bounds), the iterative scheme gives the same asymptotic bounds as the ones obtained using the entropic approach. To this day, there is no proof that the iterative scheme always converges to the "correct" value of the localized averages. Even more so, the question of when is it possible to replace the $L_2(\mu)$ ball by an empirical ball remains open.

A Concentration of measure and Rademacher averages

In this section we prove that all the L_p norms of the Rademacher averages of a class are equivalent, as long as the class is not contained in a "very small" ball.

Theorem A.1 *For every $1 < p < \infty$ there is an absolute constant c_p for which the following holds. Let F be a class of functions, set μ to be a probability measure on Ω and put $\sigma_F^2 = \sup_{f \in F} \mathbb{E}_\mu f^2$. If n satisfies that $\sigma_F^2 \geq 1/n$ then*

$$c_p \big(\mathbb{E} \sup_{f \in F} \big| \sum_{i=1}^n \varepsilon_i f(X_i) \big|^p \big)^{\frac{1}{p}} \leq \mathbb{E} \sup_{f \in F} \big| \sum_{i=1}^n \varepsilon_i f(X_i) \big| \leq \big(\mathbb{E} \sup_{f \in F} \big| \sum_{i=1}^n \varepsilon_i f(X_i) \big|^p \big)^{\frac{1}{p}},$$

where $(X_i)_{i=1}^n$ are independent random variables distributed according to μ and the expectation is taken with respect to the product measure associated with the Rademacher variables and the variables X_i.

The proof of this theorem is based on the fact that $\sup_{f \in F} |\sum_{i=1}^n \varepsilon_i f(X_i)|$ is highly concentrated around its mean value, with an exponential tail. The first step in the proof is to show that if one can establish such an exponential tail for a class of functions, then all the L_p norms are equivalent on the class. In fact, we prove a little more:

Lemma A.2 *Let G be a class of nonnegative functions which satisfies that there is some absolute constant c_0 such that for every $g \in G$ and every integer m,*

$$Pr\big\{ |g - \mathbb{E}g| \geq m\mathbb{E}g \big\} \leq 2e^{-c_0 m}.$$

Then, for every $0 < p < \infty$ there are constants c_p and C_p which depend only on p and c_0, such that for every $g \in G$,

$$c_p (\mathbb{E}g^p)^{\frac{1}{p}} \leq \mathbb{E}g \leq C_p (\mathbb{E}g^p)^{\frac{1}{p}}.$$

Proof: Fix some $0 < p < \infty$ and $g \in G$, and set $a = \mathbb{E}g$. Clearly,

$$\mathbb{E}g^p = \mathbb{E}g^p \chi_{\{g < a\}} + \sum_{m=0}^{\infty} \mathbb{E}g^p \chi_{\{(m+1)a \leq g \leq (m+2)a\}}.$$

By the exponential tail of g, $Pr\big\{ g \geq (m+1)a \big\} \leq 2e^{-c_0 m}$, and thus

$$\mathbb{E}g^p \leq a^p + 2a^p \sum_{m=0}^{\infty} (m+2)^p e^{-c_0 m},$$

proving that $c_p (\mathbb{E}g^p)^{1/p} \leq \mathbb{E}g$.

To prove the upper bound, set $h_m = \mathbb{E} g\chi_{\{g \geq ma\}}$. We will show that there is a constant $C \geq 1$ which depends only on c_0, with the property that for every $m \geq C$, $h_m \leq (\mathbb{E}g)/2$. Indeed,

$$h_m = \sum_{n=m}^{\infty} \mathbb{E} g\chi_{\{na \leq g < (n+1)a\}} \leq 2a \sum_{n=m}^{\infty} (n+1)e^{-c_0 n},$$

which is a tail of a converging series that does not depend on the choice of g. Thus, for a sufficiently large m our assertion holds.

Set $A = \{g \leq a/4\}$, and observe that

$$\frac{a}{2} \leq \mathbb{E} g\chi_{\{g \leq Ca\}} = \mathbb{E} g\chi_A + \mathbb{E} g\chi_{\{a/4 < g \leq Ca\}} \leq \frac{a}{4} Pr(A) + Ca(1 - Pr(A)).$$

It follows that $Pr(A^c) \geq 1/(4C-1)$ and thus,

$$\mathbb{E} g^p \geq \mathbb{E} g^p \chi_{A^c} \geq \big(\frac{a}{4}\big)^p \cdot \frac{1}{4C-1} = C_p a^p,$$

as claimed.

■

Before we continue with our discussion, let us observe that the exponential tail assumption can be slightly relaxed. In fact, all that we need is that the probability that g is much larger than its expectation must decay rapidly, uniformly in g.

Now, we can show that for any class of functions F, $R_n(F)$ may be bounded from below by σ_F.

Lemma A.3 *There is an absolute constant c such that for any class $F \subset B\big(L_\infty(\Omega)\big)$, $R_n(F) \geq c\sigma_F$, provided that $\sigma_F^2 > 1/n$.*

Proof: By the assumption on σ_F, there is some $f \in F$ for which $\sigma_f^2 = \mathbb{E}_\mu f^2 \geq 1/n$. Applying the Kahane-Khintchine's inequality, there is an absolute constant c such that for every $x_1, ..., x_n$

$$\mathbb{E}_\varepsilon \sup_{f \in F} \Big|\sum_{i=1}^n \varepsilon_i f(x_i)\Big| \geq c\Big(\mathbb{E}_\varepsilon \sup_{f \in F} \Big|\sum_{i=1}^n \varepsilon_i f(x_i)\Big|^2\Big)^{\frac{1}{2}} \geq c\Big(\sum_{i=1}^n f^2(x_i)\Big)^{\frac{1}{2}}$$

(in fact $c = 1/\sqrt{2}$ will suffice, as shown in [14]). Hence, $R_n(F) \geq c\mathbb{E}_\mu\big(n^{-1}\sum_{i=1}^n f^2(X_i)\big)^{\frac{1}{2}}$.

Define $g(X_1, ..., X_n) = n^{-1}\sum_{i=1}^n f^2(X_i)$ and since f is bounded by 1 then $\mathbb{E}g^2 \leq \sigma_f^2$. By Bernstein's inequality (theorem 2.20) and selecting $x = nm\mathbb{E}_\mu g$ for some integer m,

$$Pr\big\{|g - \mathbb{E}_\mu g| \geq m\mathbb{E}_\mu g\big\} \leq 2e^{-c\frac{n^2 m^2 (\mathbb{E}_\mu g)^2}{\sigma_f^2 n + nm\mathbb{E}_\mu g}}.$$

But since $\mathbb{E}_\mu g = \sigma_f^2$ then the exponent is of the order of $nm\sigma_f^2$, and because $n\sigma_f^2 \geq 1$ then there is an absolute constant c such that

$$Pr\big\{|g - \mathbb{E}_\mu g| \geq m\mathbb{E}_\mu g\big\} \leq 2e^{-cm}.$$

Using the previous lemma for $p = 1/2$ it follows that there are absolute constants c and C such that $c(\mathbb{E}_\mu g^{1/2})^2 \leq \mathbb{E}_\mu g \leq C(\mathbb{E}_\mu g^{1/2})^2$. Thus,

$$(\mathbb{E}_\mu g^{\frac{1}{2}}) \geq c(\mathbb{E}_\mu g)^{\frac{1}{2}} = c\big(\frac{1}{n}\mathbb{E}_\mu \sum_{i=1}^n f^2(X_i)\big)^{\frac{1}{2}} = c\sigma_f,$$

as claimed.

■

proof of theorem A.1: First, note that the upper bound holds, by applying Hölder's inequality. As for the lower bound, denote by $\mathbb{E}$ the expectation with respect to the product measure $\nu^n = (\varepsilon \otimes \mu)^n$ and set $H = n^{-1/2} \sup_{f\in F} |\sum_{i=1}^n \varepsilon_i f(X_i)|$. Instead of the applying Bernstein's inequality, we will use its functional version (2.6), for the random variable

$$Z = \sup_{f\in F}\big|\sum_{i=1}^n \varepsilon_i f(X_i) - \mathbb{E}\sum_{i=1}^n \varepsilon_i f(X_i)\big| = \sqrt{n}H.$$

Using the notation of theorem 2.21, $\sigma^2 = n\sigma_F^2$, and with probability larger than $1 - e^{-x}$,

$$\frac{1}{\sqrt{n}}\sup_{f\in F}\big|\sum_{i=1}^n \varepsilon_i f(X_i)\big| \leq 2\mathbb{E}\frac{1}{\sqrt{n}}\sup_{f\in F}\big|\sum_{i=1}^n \varepsilon_i f(X_i)\big| + C\big(\sigma_F\sqrt{x} + \frac{x}{\sqrt{n}}\big),$$

for some absolute constant C. By our assumption, $\sigma_F \geq 1/\sqrt{n}$, and by lemma A.3, $\sigma_F \leq n^{-1/2}\mathbb{E}\sup_{f\in F}|\sum_{i=1}^n \varepsilon_i f(X_i)|$. Thus, selecting $x = m$ for some integer m, it follows that there is an absolute constant C such that with probability larger that $1 - e^{-m}$, $H \leq Cm\mathbb{E}H$. Hence,

$$Pr\big\{H \geq m\mathbb{E}H\big\} \leq e^{-cm},$$

for an appropriate absolute constant c. Using the same argument as in lemma A.2, it follows that all the L_p norms of H are equivalent, which proves our assertion.

■

References

[1] M.Anthony, P.L. Bartlett *Neural Network Learning: Theoretical Foundations*, Cambridge University Press, 1999.

[2] N. Alon, S. Ben–David, N. Cesa–Bianchi, D. Haussler: Scale sensitive dimensions, uniform convergence and learnability, J. of ACM 44 (4) 615–631, 1997.

[3] O. Bousquet: A Bennett concentration inequality and its application to suprema of empirical processes, preprint.

[4] L. Devroye, L. Györfi, G. Lugosi: *A probabilistic theory of pattern recognition*, Springer, 1996.

[5] R.M. Dudley: *Real analysis and Probability*, Chapman and Hall, 1993.

[6] R.M. Dudley: The sizes of compact subsets of Hilbert space and continuity of Gaussian processes, J. of Functional Analysis 1, 290-330, 1967.

[7] R.M. Dudley: Central limit theorems for empirical measures, Annals of Probability 6(6), 899-929, 1978.

[8] R.M. Dudley: *Uniform Central Limit Theorems* Cambridge Studies in Advanced Mathematics 63, Cambridge University Press 1999.

[9] N. Dunford, J.T. Scwartz: *Linear Operators, part I*, Wiley 1957.

[10] E. Giné and J. Zinn, "Some limit theorems for empirical processes", *Annals of Probability*, 12(4), 929–989, 1984.

[11] D. Haussler: Sphere packing numbers for subsets of Boolean n-cube with bounded Vapnik-Chervonenkis dimension, Journal of Combinatorial Theory (A) 69, 217-232, 1995.

[12] W. Hoeffding: Probability inequalities for sums of bounded random variables, Journal of the American Statistical Association, 58, 13-30, 1963.

[13] V. Koltchinskii, D. Panchenko: Rademacher processes and bounding the risk of function learning, *High dimensional probability, II (Seattle, WA, 1999)*, 443–457, Progr. Probab., 47, Birkhauser.

[14] R. Latala, K. Oleszkiewicz: On the best constant in the Khintchine-Kahane inequality, Studia Math. 109(1), 101-104, 1994.

[15] M. Ledoux: *The concentration of measure phenomenon*, Mathematical Surveys an Monographs, Vol 89, AMS, 2001.

[16] M. Ledoux, M. Talagrand: Probability in Banach Spaces: isoperimetry and processes, Springer, 1991.

[17] W.S.Lee, P.L. Bartlett, R.C. Williamson: The Importance of Convexity in Learning with Squared Loss, IEEE Transactions on Information Theory 44 (5), 1974-1980, 1998.

[18] P. Massart: About the constants in Talagrand's concentration inequality for empirical processes, Annals of Probability, 28(2) 863-884, 2000.

[19] S. Mendelson: Rademacher averages and phase transitions in Glivenko-Cantelli class, IEEE transactions on information theory, 48(1), 251-263, 2002.

[20] S. Mendelson: Improving the sample complexity using global data, to appear IEEE transactions on information theory. Available at `www.axiom.anu.edu.au/~shahar`.

[21] S. Mendelson: Geometric parameters of kernel machines, preprint. Available at `www.axiom.anu.edu.au/~shahar`.

[22] S. Mendelson, R. Vershynin: Entropy, combinatorial dimensions and random averages, preprint. Available at `www.axiom.anu.edu.au/~shahar`.

[23] S. Mendelson, R. Vershynin: Entropy and the combinatorial dimension, preprint. Available at `www.axiom.anu.edu.au/~shahar`.

[24] S. Mendelson, R.C. Williamson: Agnostic Learning nonconvex classes of functions, preprint. Available at `www.axiom.anu.edu.au/~shahar`.

[25] V.D. Milman, G. Schechtman: *Asymptotic theory of finite dimensional normed spaces*, Lecture Notes in Mathematics 1200, Springer 1986.

[26] A. Pajor, *Sous espaces ℓ_1^n des espaces de Banach*, Hermann, Paris, 1985

[27] G. Pisier, *The volume of convex bodies and Banach space geometry*, Cambridge University Press, 1989.

[28] E. Rio: Une inegalité de Bennett pour les maxima de processus empiriques, preprint.

[29] N. Sauer: On the density of families of sets, J. Combinatorial Theorey (A), 13, 145-147, 1972.

[30] S. Shelah: A combinatorial problem: stability and orders for models and theories in infinitary languages, Pacific Journal of Mathematics, 41, 247-261, 1972.

[31] V.N. Sudakov, “Gaussian processes and measures of solid angles in Hilbert space”, *Soviet Mathematics. Doklady* 12, pp. 412-415, 1971.

[32] M. Talagrand: Type, infratype and the Elton-Pajor theorem, Inventiones Mathematicae, 107, 41-59, 1992.

[33] M. Talagrand: Sharper bounds for Gaussian and empirical processes, Annals of Probability, 22(1), 28-76, 1994.

[34] A.W. Van der Vaart, J.A. Wellner: *Weak convergence and Empirical Processes*, Springer-Verlag, 1996.

[35] V. Vapnik: *Statistical Learning Theory*, Wiley 1998

[36] A. Vidyasagar: *The Theory of learning and generalization* Springer-Verlag, 1996.

[37] V. Vapnik, A. Chervonenkis: Necessary and sufficient conditions for uniform convergence of means to mathematical expectations, Theory Prob. Applic. 26(3), 532-553, 1971